ABSTRACT

Title of dissertation: AN INVESTIGATION OF THE RELATIONSHIP BETWEEN AUTOMATED MACHINE TRANSLATION EVALUATION METRICS AND USER PERFORMANCE ON AN INFORMATION EXTRACTION TASK

Calandra Rilette Tate, Doctor of Philosophy, 2007

Dissertation directed by: Professor Eric V. Slud
Department of Mathematics
& co-directed by: Professor Bonnie J. Dorr
Department of Computer Science

This dissertation applies nonparametric statistical techniques to Machine Translation (MT) Evaluation using data from a MT Evaluation experiment conducted through a joint Army Research Laboratory (ARL) and Center for the Advanced Study of Language (CASL) project. In particular, the relationship between human task performance on an information extraction task with translated documents and well-known automated translation evaluation metric scores for those documents is studied. Findings from a correlation analysis of the connection between autometrics and task-based metrics are presented and contrasted with current strategies for evaluating translations. A novel idea for assessing partial rank correlation within the presence of grouping factors is also introduced. Lastly, this dissertation presents a framework for task-based machine translation (MT) evaluation and predictive modeling of task responses that gives new information about the relative predic-

tive strengths of the different autometrics (and re-coded variants of them) within the statistical Generalized Linear Models developed in analyses of the Information Extraction Task data.

This work shows that current autometrics are inadequate with respect to the prediction of task performance but, near adequacy can be accomplished through the use of re-coded autometrics in a logistic regression setting. As a result, a class of automated metrics that are best suitable for predicting performance is established and suggestions are offered about how to utilize metrics to supplement expensive and time-consuming experiments with human participants. Now users can begin to tie the intrinsic automated metrics to the extrinsic metrics for task they perform. The bottom line is that there is a need to average away MT dependence (averaged metrics perform better in overall predictions than original autometrics). Moreover, combinations of recoded metrics performed better than any individual metric. Ultimately, MT evaluation methodology is extended to create new metrics specially relevant to task-based comparisons. A formal method to establish that differences among metrics as predictors are strong enough not to be due by chance remains as future work.

Given the lack of connection in the field of MT Evaluation between task utility and the interpretation of automated evaluation metrics, as well as the absence of solid statistical reasoning in evaluating MT, there is a need to bring innovative and interdisciplinary analytical techniques to this problem. Because there are no papers in the MT evaluation literature that have done statistical modeling before or that have linked automated metrics with how well MT supports human tasks, this work

is unique and has high potential for benefiting the Machine Translation research community.

An Investigation Of The Relationship Between Automated Machine Translation Evaluation Metrics and User Performance on an Information Extraction Task

by

Calandra Rilette Tate

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2007

Advisory Commmittee:

Professor Eric V. Slud, Chairman/Advisor
Professor Bonnie J. Dorr, Co-Advisor
Professor Paul Smith
Professor C. Mitchell Dayton
Dr. Clare R. Voss

DEDICATION

This work is dedicated to two of my greatest influencers—there since the beginning, but unfortunately unable to witness their seed flourish:

my great-grandmother
Christine Domineck Thompson
October 22, 1906 - January 6, 2001

...Thanks for beginning my charity at home;
I promise to continue to spread it abroad greatly

my dad
Riley Jerome Tate
December 17, 1957 - October 2, 2000

...How I'd love love love to dance with my father again

ACKNOWLEDGMENTS

"For today and its blessings, I owe the world an attitude of gratitude."
˜Clarence E. Hodges

There is a long list of people in my world to whom I am forever grateful for making this accomplishment possible. Let me start by giving praise to God for His grace and mercy in the manifestation of this blessing. Next, I would like to acknowledge my academic supporters. I must thank the world-class faculty and administrative professionals of the Department of Mathematics at the University of Maryland. Innumerable thanks to my advisor, Dr. Eric V. Slud for his unwavering support of, investment in, and dedication to me and my work. I could not have asked for a more caring, respectful, fair, generous, accessible or diligent teacher-mentor. Thanks to Dr. Bonnie J. Dorr for her great expertise, guidance, enthusiasm, support and co-advising on this work (I look forward to the reunion gatherings at your place!). Lastly, I would like to thank the rest of my committee: Dr. Clare R. Voss for her initial vision in having me explore this research area. I appreciate her taking me under her wing and nurturing my development as a researcher both professionally and personally. Dr. Paul Smith—whose specific suggestions account for my pursuit of document level random effects in my modeling—and Dr. C. Mitchell Dayton both demonstrated flexibility, patience, and a desire to serve on my committee; and, they each provided very useful feedback, insight, and comments on my work.

I want to thank my family who has always been my biggest source of support and encouragement: my mother, Desireè Thompson Taylor, has been there since the beginning...directing, guiding, supporting me, and 'butting in' in her unique way; my grandmothers Lou C. Tate (Dea Lou) and Essie B. Bumpus (Gram), for their constant prayers, positiveness, and encouragement; my not-so-little brothers Jeremy, Tywone, Arsenio and my favorite nephew Jamyre, for giving me a reason to excel; and all those in my entire Tate and Thompson family—too big to name—but all of whom have been tremendously supportive and on whose shoulders I truly stand. Also, I would be remiss if I did not mention my first mentors and role-models: my cousin Dr. Terri Major-Kincade, to whom I am always compared to; thanks for giving me something to strive for and helping me to decipher early on that, although I may become a doctor one day, it certainly wouldn't be a medical doctor! And, Mrs. Terry Taylor Keller provided a safe shelter and warm support during my idle spare time as a youth when I could have done so many other things. Thanks for helping me explore my interest in math since middle school through learning to work the cash register and handle transactions in your store.

I also want to thank all of my wonderful group of friends from grade school, college, and beyond with whom I continue to form strong relationships with today: especially my confidantes Dacia Odom, Dr. Trimiko Melancon, Joycelyn Wilson, and Dr. Angela Grant for your listening ears every step of the way; thanks so much for having my back! I want to acknowledge the wide net of support from my fellow AAMath folk from Maryland, both the alumni and those still in the program, who have made this experience possible and memorable. Furthermore,

I thank the members of the numerous organizations that kept me balanced and busy throughout this process: The Louisiana Network of Washington DC and their constant flava' from home; The Usher Board of the People's Community Baptist Church (especially my youth ushers); my talented sisters of SISTERMENTORS, directed by Dr. Shireen Lewis, and especially my accountability partner—the soon-to-be Dr. Tisha Lewis; the ladies of Theta Omega Omega Chapter of Alpha Kappa Alpha Sorority, Incorporated, especially my regular dinner buddies whose routine meetings kept me going—Sorors Frances Frost, Michelle Rush, and Rae Sinanan; and last but not least, I am forever indebted to my supervisors and colleagues at the U.S. Army Research Laboratory.

Lastly, this process would not have been as bearable as it was without the endearing support of my study-partner-turned-best-friend Eulus Samuel Moore III. Through all the toils and strains of this process, I have learned and gained so much and feel doubly fortunate to leave not only with a degree but a mate for life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

List of Abbreviations

AIC	Akaike Information Criterion
ALPAC	Automatic Language Processing Advisory Committee
ARL	Army Research Laboratory
BLEU	Bilingual Evaluation Understudy
CASL	Center for Advanced Study of Language
CDER	Cover Disjoint Error Rate
DARPA	Defense Advanced Research Projects Agency
FEMTI	Framework for Machine Translation Evaluation
GALE	Global Autonomous Language Exploitation program
GTM	General Text Matcher
GLM	Generalized Linear Model
GLMM	Generalized Linear Mixed Effect Models
HTER	TER with human targeted references
LRT	Likelihood Ratio Test
METEOR	Metric for Evaluation of Translation with Explicit word Ordering
MT	Machine Translation
MOE	Measure of Effectiveness
MOP	Measure of Performance
MSE	Mean Square Error
NIST	National Institute of Standards and Technology
RMSE	Root Mean Square Error
SSE	Sum of Squared Error
TER	Translation Edit Rate
WH-type	Who, When, Where category type

Chapter 1

Introduction

Continual changes in conditions in the world around us produce a high demand for foreign language understanding. Particularly in the wake of the September 11 attacks, the Department of Defense has recognized military needs in document processing and evaluation. Translating texts from one language to another is quite a complex task. However, many systems developed throughout the years by researchers and industry that have attained some success in achieving this traditionally human-performed task with a computer. Today, there are many *Machine Translation* (MT) systems or "engines" ranging from linguistic knowledge-based engines to statistically rooted engines. There have also been hybrid systems based on a combination of the two types of engines.

1.1 Motivation

With the increase in production of many reputable language translation software engines, users like the US Government need to make choices as to which engines they should invest in based on which one provides the best output for their specific

tasks. Researchers have proposed several methods for quality assessment to tackle such concerns over the years with the introduction of various evaluation strategies. First, in the early 1990's, ordinal scale human subjective judgments were introduced and became the gold standard of translation quality. Eventually, a focus on faster, more intuitive means for evaluating translations stimulated the development of several novel automated algorithms starting in 2001. The main results in this arena are based on the correspondence of a system translation to predetermined correct *human reference* translations and are discussed further in Chapter 2.

1.2 The Problem

Current evaluation methods have considerably furthered the development of translation engines based on the system developer's ability to obtain a numerical estimate of the system's current capabilities. As a result, there has been the assumption among MT developers that MT engines are "good enough" to support people performing certain applications in the real world [12]. Yet, none of the current methods actually take into account the assessment based on the utility of the documents the systems produce even though, "there are no absolute standards of translation quality but only more or less appropriate translations for the purpose for which they are intended" [56]. More recently, informal reports from operational and field settings have described successful, but carefully limited, use of MT output in real-world tasks [23, 29].

Figure 1.1 shows a triangular diagram of the progression of MT Evaluation

paradigms. There has been extensive work to find the association between the most popular approaches on the top axis of the diagram. However not much, if any, effort has been devoted towards pursuing the other possible connections. Thus, although translation evaluation has evolved and become very important in the research and development of high caliber machine translation engines, an important part has remained left out—the practical assessment of documents from the user's perspective.

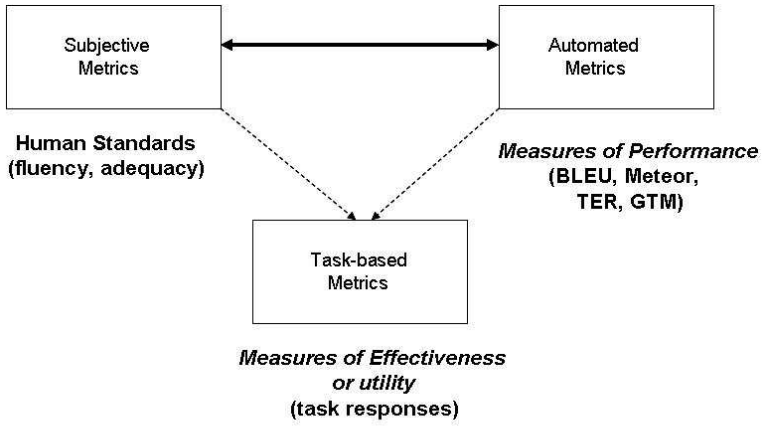
This dissertation expands the "Task-Based Metrics" node of Figure 1.1 and focuses on the connections between it and the other two nodes. Work on this problem makes use of data from an experiment where I participated in both the design and analysis, in conjunction with the Army Research Lab (ARL) and sponsored by the Center for Advanced Study for Language (CASL) at the University of Maryland [68]. The study, described in detail in Chapter 3, was designed to assess the translation output of three different MT systems and the performance of multi-level translation analysts on a Who, Where, When information extraction task using translated documents produced by these machines.

1.3 Research Questions

This study builds upon existing automated MT Evaluation metrics to connect them to interpretable task-based metrics. Specifically this work examines the following research questions:

- What is the nature of relationship between extraction task performance and automated translation evaluation metrics?

Figure 1.1: Triangle of main MT evaluation paradigms. The bold line represents the work of past efforts. The dashed line on the right represents the research explored in this dissertation. The dashed line on the left represents possibilities yet to be explored in detail.



- Do autometric scores predict user performance on an extraction task? Specifically, are there certain metrics that do a better job in predicting task performance?

- From a methodological standpoint, how can the true degree of bivariate correlation between variables of interest (metrics and performance measures) be summarized in the presence of important classifying variables (MT and WH-type)?

1.4 Statistical Overview

The data collected from the extraction task is used to build upon the work already accomplished by established automated metrics, referred to as *autometrics* in the remainder of this document, such as: BLEU [52], GTM [50], METEOR [38],

and TER [58]. These autometrics are computed for each translated document and serve as data values which are analyzed jointly with task performance rates. The relationship between task performance results and automated metrics is studied through an exploration of various aspects of correlation between the metrics and subject responses. Initial steps in this investigation involve: (1) reviewing the current correlation analysis methodologies for comparing evaluation metrics; and (2) addressing the crude results obtained in studies that use correlations to compare task response metrics with automated metrics.

This dissertation establishes that there is a positive monotonic relationship between these variables in the data; however, relationship in the presence of grouping variables is shown to be weaker than original assumed within the data studied. This demonstrated weakness motivates the need for extending beyond simple correlation for the purpose of utilizing autometrics for utility assessments of documents. Partial rank correlation, which will be discussed in Chapter 4, serves as a tool for documenting the reality of within-MT-group relationships which could be very different across groups. The large data sample distribution of this correlation statistic is approximated and then permutational methods and simulation are used to demonstrate the implications of the grouping factors associated with the cross-classification of documents into particular WH-types or MT systems. These grouping factors are explained in more detail in Chapter 5.

Through logistic regression [1, 30], the utility of a translation given specific document characteristics (such as the translation quality score) is estimated to produce statistical models for predicting user responses. By using statistical modelling,

it is possible to detect coefficients within logistic regression models that are *statistically significantly* different from 0. This indicates whether specific predictor variables have an influence on the responses collected in the experiment. Several logistic models are fitted to these cross-classified data, using correct subject matching as response variables and incorporating document and machine effects as predictors for the probability of correct matching.

Chapter 6 summarizes the specification, fitting, and interpretation of generalized linear models describing the dependence of subject performance on the extraction task on autometrics and other document features. Best fitting fixed effect models show that autometrics are useful in distinguishing task-based performance of MT engines and under specific response criteria, certain MT engines do outperform others on subject responses for the extraction task. The consequences of such models, their effectiveness, statistical adequacy, and limitations as a predictive tool are addressed. The modelling results are analyzed to determine which autometric or class of autometrics is more useful in predicting document utility. Additionally, this dissertation outlines how interpretations of the models will aid us in future testing and evaluation.

1.5 Significance of this Study

Given the lack of connection in the field of MT Evaluation between task utility and the interpretation of autometrics, as well as the absence of solid statistical reasoning in evaluating MT, there is a need to bring innovative and interdisciplinary

analytical techniques to this problem. Because there are no papers in the MT evaluation literature that have done statistical modelling before or that have linked autometrics with how well MT supports human tasks, this work is unique and has high potential for benefiting the Machine Translation research community. We begin here to address the pertinent needs of users, such as the Department of Defense, in document processing and evaluation.

Although reliable results are best obtained through task-based experiments and evaluation methods that analyze real-world task performance using multiple MT engines, task-based experiments can be quite time-consuming and labor intensive. As noted in [16], resource considerations such as these have forced the field to rely heavily on automated metrics. Thus, it is crucial in any evaluation to determine how well results with these metrics compare to the results found in task-based analysis. In particular, we want to know whether there is a relationship between these popular, strictly text-based metrics and the end-to-end (machine and user) effectiveness metrics of concern to real users. This dissertation proposes a method for evaluating the performance of a translation system by analyzing how accurately subjects perform on a task that incorporates output of the translation system.

This work motivates the need to extend beyond limited descriptive statistical analysis and utilize statistical models to develop other uses of these autometrics for a more user-centered evaluation. It is the goal that users, as well as researchers, could benefit from a practical, *task-based* measure of the operational capability description of MT performance in terms of autometrics using the predictive modelling strategies of this dissertation.

This dissertation directly yields the following contributions:

Applied Statistical Tools and Methodology

- Methodological results on characteristics of bivariate and partial rank correlation analysis, coupled with permutation tests of significance

Application to Machine Translation Evaluation

- A first, in-depth, user-centered focus of translation evaluation using autometrics and an innovative analysis of data through statistical modelling techniques to make use of autometrics as tools for assessing translated documents
- Assessment of *document-level* correspondence between autometrics and hit rate, and the refinement of this correspondence to apply within MT by WH groups
- Identification of autometrics that offer best predictions of user task performance in comparison to competing metrics

1.6 Chapter Overview

The remainder of this dissertation is organized as follows:

- Chapter 2 gives an overview of the existing literature and relevant work on Machine Translation Evaluation. Background information is presented on the common automated evaluation metrics that are used in this research, as well

as other evaluations and research involving humans-in-the-loop and task based MT evaluation procedures.

- Chapter 3 provides an overview of the unique, large-scale task-based experiment which yielded the data used in this work. The data collection procedures, experimental design, and answer set creation are discussed.

- Chapter 4 reviews the correlation analysis methodology that is currently used in comparing evaluation metrics. Correlations between task-based metrics and automated metrics are presented and an extension of this analysis is provided that involves partial correlation results. This chapter presents the primary motivation for extending beyond simple correlation determining the effectiveness of automated metrics for assessing document utility.

- Chapter 5 offers a methodological view of the partial Spearman rank correlation. The asymptotic properties of two versions of this statistic are demonstrated under certain conditions and empirical simulation results are used to compare power of the statistics against different alternatives.

- Chapter 6 investigates the use of statistical modelling techniques, namely logistic regression, to acquire a sense of the predictive impact of four known automated metrics—BLEU, GTM, METEOR, and TER—on performance in the WH-extraction task. Each model is discussed in detailed along with a discussion of statistical adequacy as measured by goodness of fit criteria.

- Chapter 7 presents an analysis of the modelling results to determine which

metric or class of metrics is more useful in predicting document usefulness as assessed by task performance rates. Interpretations of the models are shown to enable future testing and evaluation. A summary of the results of this dissertation and present recommendations for future extensions of this research are presented.

Chapter 2

Background and Related Work

Not only has Machine Translation grown as an important research field in Natural Language Processing, but the sub-area of Machine Translation Evaluation has also become a very active area of research. Yorick Wilks [71] has often been quoted for his comment that MT Evaluation is about as studied as MT alone and that there is more discussion about evaluation of MT than MT itself. The infamous ALPAC report [4] proved that evaluation results have great potential to influence the direction of MT research in the future. Dorr [20] provides an in-depth survey of MT paradigms including evaluation. Organizations such as FEMTI, The Framework for Machine Translation Evaluation, in International Standards for Language Engineering [32] have organized a comprehensive guide of the various methods that are used to evaluate MT systems [35].[1] In the following sections, the major shifts in evaluation efforts over the years are outlined and the overarching questions that drive this dissertation work are discussed.

[1] Additional information about the FEMTI project and the resources it offers can be found at: http://www.issco.unige.ch:8080/cocoon/femti/st-home.html

2.1 Human Translation Evaluation

One bottleneck of MT Evaluation is that most strategies, to some extent, involve human labor. Whether the role of the human is to provide judgments on the quality of MT outputs, to translate documents for later comparison to MT output, or to perform tasks on translated output to obtain utility metrics; the need for humans is still there. Furthermore, perhaps because translation is a human-oriented task, it may just be the case that evaluations will always require the humans to support evaluation methods. In this section, I review the relevant practices involving humans for MT Evaluation.

2.1.1 Human Quality Judgments

Human judgments of translation quality, although subjective, have been used as the standard by which to measure any evaluation technique. The idea is that if a metric can rank system output comparably to a human's ranking of the texts, the metric is deemed to be a good marker for measuring translation quality. Thus, most evaluators show metric reliability by comparing scores to a human ranking of the same document set. The assessment method followed by human judges most often involves examining translation quality with respect to *adequacy*, *fluency*, and *comprehension*. These figures of merit were used by DARPA in its MT evaluations during the early 1990s and are each described below [70]:

Description of Judgments[2]

[2]These measures have since been adapted and refined by the Linguistic Data Consortium (LDC) for use in current NIST MT evaluations. An in-depth description of the procedure for obtaining

- Adequacy is judged by comparing each translated segment with the corresponding segment of a human-produced reference translation. The segment's score for adequacy is measured according to how well the meaning of the reference translation is conveyed in the meaning of the system's translation under evaluation. The judges are instructed to rate the degree to which items are preserved between the system and reference translation on a scale of one to five. If the meaning was absent or not conveyed properly, the score given should be one and if it was completely present, the score would be five.

- Fluency is scored independently of the source document or of any reference translation. This can be done by a monolingual speaker of English. This criterion assesses a native speaker's intuition about the proper English sentence structure of a text on a segment by segment basis. Evaluators assign a score between one and five to each unit with five denoting a perfectly structured English sentence and one denoting a sentence with extremely bad organizational structure.

- Comprehension measures the amount of information that is correctly conveyed, i.e., the degree to which a reader can find usable information in text. In the DARPA study, this evaluation was in the format of a standardized comprehension test. Questions were developed based on the human reference translations and then used for the system translation. Evaluators were instructed to base their answers only on information present in the translation.[3]

human judgments for translation can be found on their website at: http://projects.ldc.upenn.edu.

[3]This particular measure has since been referred to as *informativeness* and serves as a fidelity

2.1.2 Human Reference Translations

In addition to human *judgments* of translation outputs, humans are also needed to create a human reference set of translations for a given source language. Good human translations are considered the gold standard for translation quality and are heavily relied upon in evaluation using the automated metrics discussed in the next section. It is the belief of most of the evaluation community that translations produced by a system should be considered good only if they are close to a good human reference translation. Thus, the way to compute an automated metric score for a document set is to compare to a human produced translation of the same documents. Initial creations of human reference translations can be very labor-intensive, but once a set is created, the collection of source/reference translation pairs can be reused many times for system testing and evaluation. In practice, these parallel corpora are very valuable and highly sought after in Machine Translation research.

2.1.3 Task Based Evaluation

Task-based evaluation provides a practical way of evaluating translation by a *measure of effectiveness (MOE)* enabling users to identify what types of tasks can be performed using the output. Church and Hovy [12] gave insight into this approach by proposing exploration in MT evaluation from the standpoint of what can be gained from the "crummy MT output." In the 1990's new MT research trends

measure for determining if an assessor can find the required 'information' in a specific translation. Many recent evaluations have only focused on the two measures: Adequacy and Fluency.

emerged, furthering interest in extrinsic metrics. Task-based experiments assuming an ordering of task difficulty were proposed by users on text-handling tasks [61, 21]. The task hierarchy produced by [61] served as a basis for the task conducted in this study. MT developers [55, 43] have also conducted prior task-based experiments evaluating text and speech MT. In the case of [55] the "gisting" or categorization task was tested among experimental subjects to determine a procedure for how well subjects can get the gist of website content with and without the use of various translation tools. The work of [43] evaluated "goal accomplishment" using a speech to speech translation system. Automated metrics had not yet arrived on the scene, so both of these studies were strictly task-based.

This alternative approach to document quality evaluation, based not on an intrinsic question of what is actually in the document, but more extrinsically on what one can do with the documents, has formed the basis for this research. This notion is more forgiving of translation quality in the sense that even weak translations may still be sufficient for certain tasks. Although this method has not been explored as much as other approaches, acquiring a connection between intrinsic and extrinsic metrics for MT evaluation would give users, as well as researchers, a threshold for task performance relative to machine performance. Vanni et al. [63] have begun developing a similar connection between evaluation metrics and linguistic assessments of MT output. MT stakeholders also began funding research experiments in task-based assessment of MT engines, to address users' needs. [4]

[4]See the 2005 broad agency announcement (BAA) for the Global Autonomous Language Exploitation program (GALE) released by DARPA, a US government funding agency.

2.2 Automated Evaluation Metrics Used in This Study

Human judgment evaluations are not only highly expensive and labor intensive but also offer human-error prone subjective judgments. While taking account of the notion that good human translations are costly but yet still the standard, researchers began developing ways to compare system output to different human reference translations. The most recent wave of evaluation research put significant effort into various automated evaluation metrics that compare MT systems using *measures of performance (MOPs)*, which measure MT output accuracy. This two-part process requires construction and annotation of an evaluation corpus. Researchers must first build the collection of multiple (human created) reference translations for the MT output to be evaluated. Secondly, the method requires specific tagging of both the human reference translations and the MT output. After this preprocessing, the measures assign a score to a "candidate" or translated output based on a predetermined algorithm to compare the output and a reference document from the collection. The individual characteristics of the four automated metrics utilized in this research are described below.

2.2.1 BLEU

In 2001, IBM researchers Papineni et al. [52] proposed the Bilingual Language Evaluation Understudy (BLEU) metric,

$$BP \times \exp\left(\sum_{n=1}^{N} w_n \log\ p_n\right) \tag{2.1}$$

This metric is the most widely used of the automated metrics for MT and is a measure of "precision." Candidate translations are scored against a user-selected number of stored reference translations by counting the number of consecutive word groups of size n, or *n-grams*, that overlap between the candidate and reference. A combination of these matches (n-gram precision) using the geometric mean across different values of n and a brevity penalty for shorter machine translations produces a translated document quality score between 0 and 1. Precision is computed by the following formula:

$$p_n = \frac{\sum_i (\text{n-grams in segment } i \text{ that match in candidate \& ref translation})}{\sum_i (\text{n-grams in candidate segment } i)}$$

Uniform weights $w_n = 1\backslash N$, where $N = 4$ (the highest number of n-grams BLEU considers), treat n-gram matches of all lengths equally. However by nature of the geometric average, higher n-grams are favored. That is, the more matches of consecutive word sequences, the higher the BLEU score and vice-versa: if high n-gram matches are not found, the score is low. The brevity penalty (BP) penalizes translations that differ significantly in length from the reference translations and thus, prevents gaming that occurs by purposefully produced short translations.

As its creators believed, BLEU has accelerated the MT R&D cycle by allowing researchers to rapidly home in on effective modelling ideas. This novel automatic approach was (and still is) perceived as an immense advantage to MT evaluation, especially in view of the fact that system performance as judged by BLEU appeared to correlate highly with time consuming and expensive human quality judgments. The inventors showed that BLEU obtained correlation coefficients against monolingual

human judgments on a corpus of Chinese-to-English translations as high as 0.99, while that with the bilingual judgments was 0.96. It is important to note here that this study was on *system*-level rankings. The analysis in Chapter 4 demonstrates why this level of aggregation can be misleading, and shows results for the purpose of this dissertation at a finer level of granularity. In an effort to get a more "sensitive" metric for evaluation, Doddington (2002) at NIST introduced an alternative version of the BLEU algorithm that uses the same co-occurrence statistics approach but uses the arithmetic mean in place of the geometric mean.

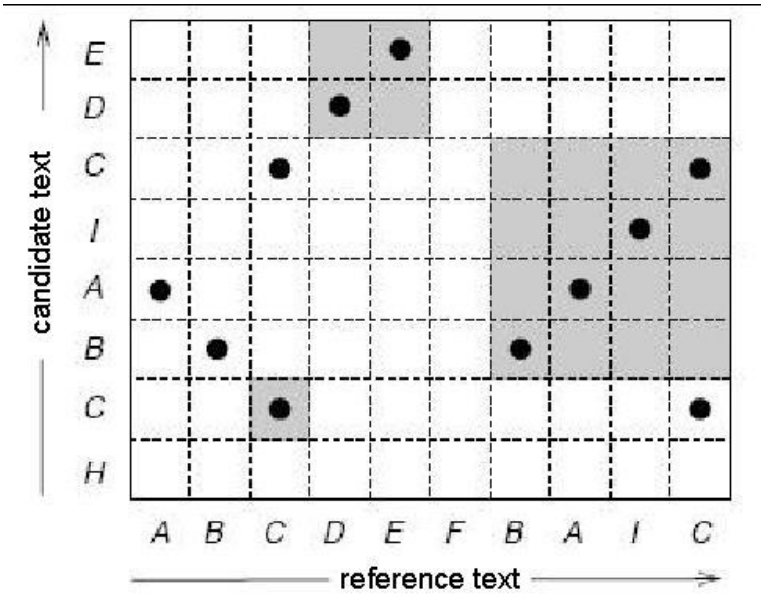
2.2.2 GTM

Turian et al. [62] established another automated approach to MT Evaluation that warrants mention, the General Text Matcher (GTM), that builds on an earlier idea by Dan Melamed [49, 50]. Candidate translations are scored with respect to a reference translation by computing similarity through the number of matching words. Unlike the BLEU method which scores translations mainly on precision, Melamed abandons the "precision only" idea altogether by using the *F-score*, a scoring function that takes the harmonic mean of precision *and* recall and provides a value between 0 and 1.

This method uses a graph theory technique called "maximum matching" to compute a maximum match size (MMS) by summing the maximum possible word length of all matching phrases between the candidate and reference translations that do not use words from any other matching. These values are computed using counts

of unigrams (hits) or aligned blocks (runs) from a bi-text grid of the candidate and reference translations. Figure 2.1, as found in [62], offers a graphical explanation of the concept. The horizontal axis is the reference text and the vertical axis is the candidate. Here there are three matches, of length 1, 2 and 4. This makes up the "maximum match" from all possible matches, calculated as $MMS(C,R) = \sqrt{(1^2+2^2+4^2)} \approx 4.6$. The square root of the sum of squares is introduced to reward longer matches and measure fluency of the translation.

Figure 2.1: Example of bitext grid to calculate MMS



Next, precision and recall are calculated by dividing the MMS by the length of the candidate ($|C|$) and the length of the reference ($|R|$), respectively:

$$Precision(C|R) = \frac{MMS(C,R)}{|C|} \qquad Recall(C|R) = \frac{MMS(C,R)}{|R|}$$

The F-score of the two measures is computed to get the final score:

$$F = \frac{2PR}{(P+R)}$$

Turian et al. assert that their F-measure is more reliable than BLEU and that it is easier to understand in terms familiar to NLP researchers.

2.2.3 METEOR

Researchers at Carnegie Mellon University introduced the Metric for Evaluation of Translation with Explicit ORdering (METEOR) for MT evaluation in 2004 [38, 7]. At the outset, developers stated that it was designed to address the following weaknesses found in the other commonly used metrics, especially BLEU:

- *The Lack of Recall*— As [50] proposed, METEOR inventors also believe that recall is a better way to measure translation quality than precision alone or the use of unclear brevity penalty measures.

- *Use of Higher Order N-grams*—The idea behind METEOR is that explicit re-ordering, done in one of its modules, better accounts for grammatical correctness than normal n-gram techniques and thus, does not use higher order n-grams in its calculation. Basically, a reordering penalty is calculated on how many chunks in the produced text need to be moved around to get the reference text.

- *Lack of Explicit Word-matching Between Translation and Reference*— METEOR tries to address the fact that incorrect matches can be obtained from n-grams because they do not require an explicit word-to-word matching by aligning candidate and reference texts.

- *Use of Geometric Averaging of N-grams*— This directly addresses BLEU's use of the geometric mean that results in zero scores for segments without

matching n-grams. [5]

METEOR heavily relies on an algorithm for finding an optimal word-to-word matching between a candidate MT translation and a human-produced reference translation for the same input sentence, and recall is a major contributor to the score. Like the other metrics, METEOR also produces normalized scores in the range of (0,1). The score is computed as follows. First, unigram precision (P) is found by counting the number of unigrams in the candidate translation that are mapped to unigrams in the reference translation and dividing by the total number of unigrams in the candidate. Likewise, unigram recall (R) is computed by dividing the same count by the total number of unigrams in the reference translation. The F-mean, a combination of the precision and recall via a harmonic-mean, is obtained as in GTM. However in this case, most of the weight is placed on recall. The resulting formula used is:

$$F = \frac{10PR}{(R + 9P)} \tag{2.2}$$

designating recall to be weighted 9 times more than precision.[6] Fluency is addressed via a direct penalty answering the question, "How fragmented is the matching of the MT output with the reference?" In other words, the *fragment count* or the number of phrases of consecutive matching words between the candidate and reference translation is sought. It is the assumption that the longer the consecutive word matches, the fewer the number of fragments. In the extreme case, where the en-

[5]This shortcoming has been addressed by NIST in their alteration of BLEU to use an arithmetic mean and by BLEU creators, themselves, with a release of a *smoothed* version of the BLEU code.

[6]The harmonic mean for values $x_1, x_2, ..., x_n$ is determined by the formula $\sum_{i=1}^{n} w_i / \sum_{i=1}^{n} (w_i / x_i)$. Equation 2.2 results since the weight is 1 for precision and 9 for recall.

tire system translation string matches the reference translation, there is only one fragment. The opposite extreme, if there are only unigram matches between the candidate/reference pair, yields as many fragments as there are unigram matches. The final METEOR score incorporates this as:

$$F * (1 - DF)$$

where the discounting factor is $DF = 0.5 * (frag^3)$ and

$$frag = \frac{\text{fragment count}}{\#\text{ unigrams matched}}.$$

The main value of this metric is that it goes behind the simple strict matching and also matches words that are simple morphological variants of each other or words that are synonyms of each other by employing the Porter stemmer and WordNet, respectively, in modules called by the user. METEOR has been shown to correlate better with human judgments than other metrics at the system level. Furthermore, METEOR investigators have specifically shown promising results (although not in comparison to other metrics) at the desirable sentence/segment level of granularity.

2.2.4 TER

Translation Error Rate (TER) [58] measures the minimum number of edits required to change a candidate output into one of the available human references. The score is normalized by the average length of the references and only uses edits recorded from the closest reference. TER uses an edit distance measure similar to word error rate to find the translation/reference pair that has the minimal number

of edits and assigns this score as the translation quality metric for the particular translation. The objective of TER is to have a repeatable human measure of fluency and meaning while providing an output that users can understand by measuring the amount of work needed to make the document both fluent and correct and simply answering the question, "What is the required number of edits for a human to *fix* a translation?"

The possible edits that TER allows include insertion, deletion, and substitution of single words as well as shifts of word phrases. A shift moves a contiguous sequence of words within the translated output to another location within the translation. All edits count as one edit, including a shift of any size or missing punctuation marks or capitalization errors. Once all the edits between the translation and (closest) reference are determined, TER is calculated using the formula:

$$TER = \frac{\text{total number of edits}}{\text{average number of reference words}} \tag{2.3}$$

The calculation can be explained in the following example provided by Snover et al. Consider the reference/hypothesis pair below, where differences between the reference and hypothesis are indicated by upper case:

REF: SAUDI ARABIA denied THIS WEEK information published in the AMERICAN new york times

HYP: THIS WEEK THE SAUDIS denied information published in the new york times

If TER is applied to this hypothesis and reference, the number of edits is 4 (1 Shift, 2 Substitutions, and 1 Insertion), giving a TER score of $\frac{4}{13} = 31\%$.

There is also a non-automatic human-in-the-loop version of TER called HTER (TER with human targeted references) that requires post-editing of system output. Here, a fluent speaker of English creates a new reference translation targeted for the particular system output by editing the translation until it is fluent and has the same meaning as the reference(s). TER is now computed including the new (human-targeted) reference. The authors have shown that HTER has the highest correlation with human judgments at sentence level. However, as the authors point out, this variation can be quite expensive. Nevertheless, HTER was recently selected as the metric of choice for the Global Autonomous Language Exploitation (GALE) research program [51].

For the purposes of this study, an altered version of TER will be used, called oTER. This measure is the original TER score multiplied by .10 and subtracted from 1.[7] This transformation was done so that TER could be directly comparable to the other evaluation metrics as they are similarity metrics and TER is an error metric. Otherwise the TER and other scores are direct opposites in magnitude on a scale of 0 to 1.

2.3 Discussion

Extrinsic, task-based evaluation of MT engines has long been of interest to the MT user community which seeks automated support tools to expedite their decision-making tasks [59]. However, although human evaluation is the ultimate

[7]The TER code I acquired produced scores ranging from 0 to 100. Later versions of this code automatically output scores between 0 and 1 so the .10 conversion may not be needed in that case.

goal, it is very time consuming and not readily re-usable. The introduction of autometrics filled a longstanding gap between MT system development and evaluation by providing researchers and users quick insight into the assessment of the quality of a certain translation. Furthermore, the creators of these metrics showed a strong correlation between the metrics and highly labor intensive and expensive human judgments. Despite these outstanding advances in MT evaluation, there are still several drawbacks to approaching MT evaluation solely using autometrics:

Drawbacks of Autometrics

- All four metrics introduced rely heavily on a human reference translation to produce a similarity score for any system translation. This, in itself, presents many difficulties. First, there is never just one way to correctly translate a document. Thus, given any foreign language text in the evaluation corpus, there will be several possible reference translations making the need for many references most desirable. Secondly, the beginning step of building the corpus is highly labor intensive and costly.

- Finding and recruiting qualified bilingual human translators is also an extremely hard task when you consider assessing a translator's language competence while taking into account that very few Americans are equally proficient in two languages. Still, one could argue that this cost is minuscule compared to the cost of hiring people to do new judgments repeatedly.

- Lastly, and perhaps most importantly, there is no solid understanding of what a specific automatic score means. How far off is a score of .35 from a score

of .50 when you are dealing with translated outputs? Is the .15 BLEU score difference really that significant? Or better yet, is it the same as the difference from .60 and .75? Likewise, is a translation quality score of .20 twice as bad as one with a score of .40? There has been no validation of this linearity of scores, and likewise, there have been no empirical results indicating how useful a translation is based on these scores.

Because of questions like these, although autometrics have become standard and have offered much insight in the evaluation community, their limits, stability, and interpretation still remain questionable. Natural questions to ask are, *How can these autometrics be employed to improve MT evaluation? Will these metrics correlate with other evaluation methods and with humans? If so, how?*

It is clear that MT evaluation research would benefit from a thorough analysis of established automated MT evaluation metrics and where they fail to measure inaccurate translation output as it relates to certain tasks. This research seeks a comprehensive autometric assessment to test and validate the described automatic metrics commonly used in MT system evaluation. This entails assessing automated measurements collectively to determine if metric scores can predict task-based results from a Name Entity Recognition task in which subjects are asked to extract proper or important names/entities representing the *Who*, *When*, and *Where*. The next chapter describes the experiment performed for this research and the data collected. This dissertation presents an alternative to intrinsic metrics by using extrinsic metrics to leverage the automated metrics in assessing the quality of MT

system output based on subject task performance. Conceivably, MT evaluation results will become more informative if a BLEU score of .20 or a GTM F-score of .65 can be correlated with specific task ability. This may get the community closer to the initial purpose of having a translation in the first place, to carry out some mission or another.

Chapter 3

Description of Experiment and Project Data

Through an experiment performed in conjunction with the Center for Advanced Study of Language (CASL) and the US Army Research Lab, I have been able to design and construct a very rich and unique dataset to use in the analysis for this research. Data was collected from a large-scale MT evaluation experiment where 59 subjects extracted who, when, and where-type essential elements of information from output generated by three types of Arabic-English MT engines. The information extraction experiment was one of three experiments performed in the study as a test of tasks of varying levels of difficulty as originally proposed by [61]. The proposed hierarchy of text handling tasks for MT output is shown in Table 3.1 with a row added for the information extraction task. This paricular task was chosen after reviewing the task hierarchy and examining the MT output of several engines. A small, prior pilot experiment to evaluate Arabic-English MT engines for document exploitation tasks indicated that subjects could extract some named entities and event participants from noisy MT output, but they could not readily identify relations within events [67]. Thus, the task was designed as an intermediate

challenge between event-level analysis and named-entity recognition.

Table 3.1: Task hierarchy by Taylor & White (1998) with extra row inserted for WH-type extraction task

Publishing	Produce technically correct document in fluent English
Gisting	Produce a summary of the document
Extraction	For documents of interest,capture specified key information
Deep Extraction	Event identification (scenarios): determine an incident type and report all pertinent information
Intermediate Extraction	Relationship identification: member-of, relative-of, boss-of
WH-Item Extraction	**Identification of: who-,when-,where-type elements**
Shallow Extraction	Named entity recognition: isolate names of people, places, organizations, dates, locations
Triage	For documents of interest, rank by importance
Detection	Find documents of interest
Filtering	Discard irrelevant documents

The next sections provide an overview of the experiment with a brief description of the document collection, the experimental design, the task and WH-type elements to be extracted, and the data collected. The full evaluation study was conducted on two days, with 30 subjects participating on day 1 and another 29 subjects on day 2. On both days the information extraction experiment, consisting of a training phase, a practice phase, and the evaluation phase took roughly two hours. The experiment was monitored by several observers and the software, run from an off-site server, was controlled by an administrator who monitored the subjects' progress online in real-time during the experiment.

3.1 WH-Extraction Task

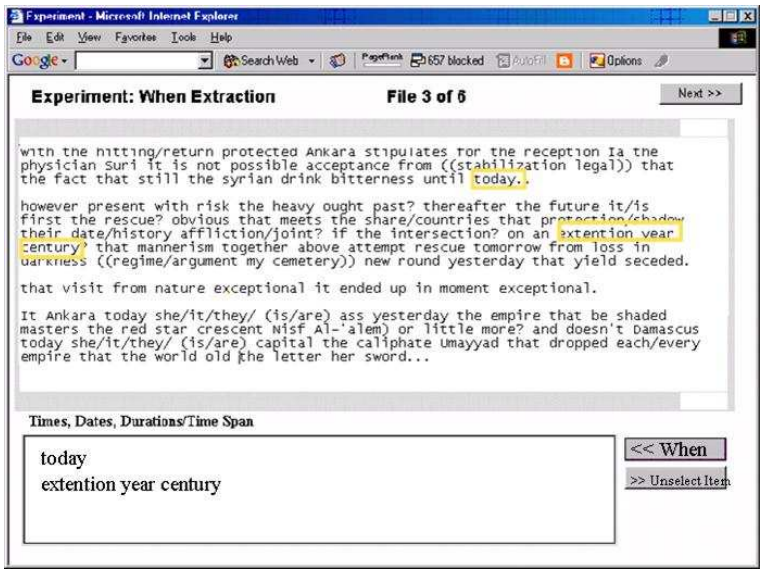
Subjects were trained to identify *who-*, *when-*, and *where*-type elements in MT output (see Table 3.2 for a description). Subjects received hard-copy pages during the *training phase* with definitions and example WH-type items that they could refer to at any phase in the experiment.

Table 3.2: Description of WH-type items in extraction task

Who-type: people, roles, organizations, companies, groups of people, and the government of a country
Where-type: geographic regions, facilities, buildings, landmarks, spatial relations, distances, and paths
When-type: dates, times, duration or frequency in time, including proper names for days and common nouns referring to time periods

During the *practice phase* with 9 documents (1 original English and 2 translated English for each of the 3 WH-types), there was no spoken instruction but all subjects practiced the task on the same sequence of documents and received the same feedback with correct responses and brief descriptions via their browsers, following their responses. The experiment software was designed to enable subjects to view all documents via the browser at their individual computer workstations and simply to click over text they selected as WH-items that would then appear in an answer box below (see Figure 3.1 for a mock-up of the screenshot subjects saw). Details of the entire experiment can be found in [69].

Figure 3.1: Mock-up of screenshot for when-extraction



3.2 Document Collection

A collection of Arabic news documents taken from ten websites was created in December 2003. Full articles were trimmed from the bottom up to be roughly comparable in size and fit fully within the software display window after translation, so that subjects would not need to use a scrollbar to see any portion of the text.

For each of the three WH-item types, native Arabic speakers identified six different trimmed documents with between six and ten WH-items of that type in the text. The documentation of these WH-items in this 18 Arabic document collection, established the "ground truth" (GT) items that encompass the experiment's answer set and determines the correct and incorrect response variables analyzed in later chapters. All 18 Arabic source documents were then run through each of three MT engines, yielding a full collection of 54 translated documents. This set of documents

forms the base collection upon which all of the automated metrics and parameters for analysis are computed.

3.3 Selection of MT Systems

Three distinct types of MT engines were selected as representatives of three development models, varying in required funding, time, and linguistic resources:[1]

- **MT-1:** a rule-based engine with hand-crafted lexicons and symbolic linguistic processing components (morphological analyzer, parser)
- **MT-2:** a statistical engine trained on large quantities of monolingual and parallel Arabic-English texts, but with no traditional symbolic linguistic processing components
- **MT-3:** a substitution-based engine that relies entirely on a pattern-matching algorithm with a lexicon and morphological analyzer to translate matched strings into English phrases, replacing the former with the latter but leaving the original Arabic word order unchanged except as occurs locally within the substituted phrases.

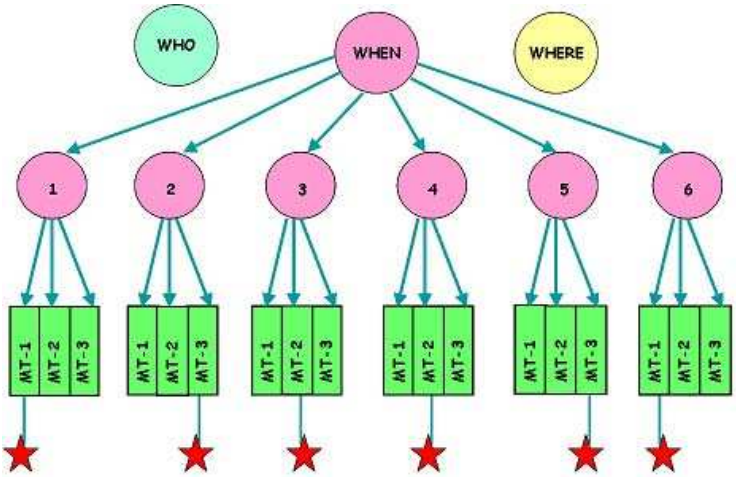
3.4 Experiment Design

The design for the experiment assumed 60 subjects, 3 MT systems, and 18 Arabic documents to be translated by each system. A diagram of the experimental

[1]The versions of the systems under study date back to 2003. The developers of the selected engines provided their most recently released version in November 2003, in preparation for this experiment conducted in January 2004.

design is pictured in Figure 3.2. I used a balanced incomplete block design [8], [46], meaning that all categories (WH and MT) were viewed an equal amount of times and not all documents were viewed by any one subject. Each subject was assigned a pre-arranged, randomized sequence of 18 documents out of the full pool of 54 machine-translated documents.

Figure 3.2: Diagram of the Experiment Design: for each WH-type, there were 6 documents of each type called Reps (represented by the circles labeled 1-6) each translated by the 3 MT systems (represented by the vertical rectangles) and then distributed to subjects according to certain project constraints. The full expansion is shown here for *when*; the *who* and *where* types have a similar structure.



The sequences were constructed as follows. First one super-block (a block containing a complete replicate of all the documents under study) was filled with three viewing sequences that included the entire 18 document × 3 MT system (54 total machine-translated documents) set, as shown in Table 3.3. Each translated document was classified and indexed by a unique MT identifier (MT-1, MT-2, or MT-3), WH identifier (WHEN,WHERE, WHO), and document identifier or Rep

Table 3.3: One super-block viewing sequence for 54 machine-translated document set containing distributed to 3 subjects x, y, z. Each document was denoted by the triple: MT system(1, 2, or 3), WH-type (WHO, WHERE, or WHEN), and document number or Rep (1, 2, 3, 4, 5, or 6).

Subject x			**Subject y**			**Subject z**		
MT	WH-type	Rep	MT	WH-type	Rep	MT	WH-type	R
1	WHO	1	3	WHO	1	2	WHO	1
3	WHEN	1	1	WHEN	1	2	WHEN	1
1	WHERE	1	2	WHERE	1	3	WHERE	1
3	WHO	2	1	WHO	2	2	WHO	2
2	WHEN	2	3	WHEN	2	1	WHEN	2
2	WHERE	2	3	WHERE	2	1	WHERE	2
1	WHEN	3	2	WHEN	3	3	WHEN	3
3	WHERE	3	1	WHERE	3	2	WHERE	3
2	WHO	3	3	WHO	3	1	WHO	3
3	WHEN	4	1	WHEN	4	2	WHEN	4
3	WHO	4	2	WHO	4	1	WHO	4
3	WHERE	4	1	WHERE	4	2	WHERE	4
1	WHEN	5	2	WHEN	5	3	WHEN	5
2	WHO	5	1	WHO	5	3	WHO	5
1	WHERE	5	2	WHERE	5	3	WHERE	5
2	WHEN	6	3	WHEN	6	1	WHEN	6
1	WHO	6	2	WHO	6	3	WHO	6
2	WHERE	6	3	WHERE	6	1	WHERE	6

(1-6). To ensure a balance of the number of MT and WH-type viewings per subject across the document collection, the following set of four constraints were placed on the experimental design:

1. No subject saw the same document more than once, i.e., as translated by more than one MT system.

2. Each subject viewed translated documents from each system an equal number of times.

3. Each subject viewed translated documents from each of the three WH-type

an equal number of times.

4. Each subject viewed the documents of each WH-type one after another, without shifting back and forth among the three WH-type groups. The order of these WH-types were randomized for each subject.

After filling the super-block to meet these constraints, the next steps were to randomize the documents by permuting the elements within each block while preserving the WH-type grouping of constraint (4) within each viewing sequence of the super-block, and then including enough replications for all subjects. Randomization was done to prevent bias. Two super-blocks were randomized individually. Then the resulting 6-block randomization was replicated 10 times for the proposed 60-person full experiment. Furthermore, since all documents of a particular type (such as WH) were viewed consecutively, permuting the order for the 3 types was considered as an added precaution against ordering effects.

3.5 Data Collected

The complete experiment, with 59 subjects each viewing 18 translated documents (arranged to include 6 documents from each of the 3 MT engines and 6 from each of the 3 WH-types), yielded a total of 1062 *cases* (in each of which one subject extracted WH-items from one document translated by one MT engine). The study had 59 subjects because one subject of the originally expected 60 did not show. Additionally, with one server-client connection crash on day 2, two instances of translated documents viewed and marked by subjects could not be processed. This

left 1060 cases for analysis. With more than 100 WH-items total in the 54 translated collection, well over 10,000 subject-extracted items were collected for evaluation.

3.6 Answer Set

The creation and validation of the WH-item answer set against which the subjects responses were scored, was constructed in three stages.[2]First, *ground-truth WH-items* (GT) were identified in the original Arabic documents by a native Arabic speaker who placed them in an inventory spreadsheet, with one item per row. The original documents were also fully translated by each of four human translators working on the project, providing reference translations for comparing the task-based results with automated metric scores in later chapters.

In the second stage, the lead translator and one professionally trained native-English linguist worked together with the resulting reference translations and the GT-annotated original documents. They identified the *reference-truth WH-items* (RT) in the reference translations and placed each of these alongside the corresponding GT item in the inventory spreadsheet.

In the final stage, six individuals independently examined pairs of the reference translations and MT outputs side-by-side, and recorded into their separate inventory spreadsheets the *omniscient-truth WH-items* (OT) in the MT outputs, that is, those strings of words found to correspond to the RT items.

The annotation of the GT, RT, and OT items across the Arabic texts, the reference translations, and the machine-translated texts, respectively, produced two

[2]For a full explanation of the details, see [65].

types of extraction-task answers used in the subsequent analysis:[3]

Set of RT items

- Independent of the MT engines
- Defined in one-to-one correspondence with the GT items[4]
- Used as denominators to calculate recall metrics (see next section)

Set of OT items

- MT engine-specific
- Defined and annotated in correspondence to the RT items[5]
- Compared against the subjects' responses to determine correctness[6]

For each Arabic document in the document collection, and for each machine-translated version of that document, the set of RT items was fixed. By contrast, for each machine-translated version of each Arabic document, the set of OT items could, and often did, vary both in content of the translated item and in number of translated items (because some were lost in translation).

3.7 Task Metrics

Three types of task metrics or *event counts* were tallied for each of the 1060 cases in the evaluation, by comparing and identifying all of each subject's responses

[3]The answers described here should be distinguished as established in documents before the experiment from the responses that subjects provide in the actual experiment.

[4]These are the English equivalents of the WH-items in the Arabic texts.

[5]These are the machine-translated equivalents of the WH-items in the Arabic texts.

[6]This is the case because the OT items are the content of the WH-items as made available to the subjects by each MT engine.

against all of the (OT) answer items in the translated document for that case. Each response is computed as follows:

- a *correct response* if a response fully matched an answer item, by covering all open class words, but possibly under- or over-extending with a determiner or other closed class word not crucial to the meaning of the WH-item
- an *incorrect response* if a response did not match any answer item in the translated document
- a *non-response* if no part of an answer in the translated document was captured in any of the subject's responses.

This dissertation focuses on the correct responses and the *correct response rate* derived from the proportion of fully correct responses out of the RT total (the total number of RT items in the reference translations). It is this collection of performance responses for documents that is compared with autometrics. The RT items are chosen as the response rate denominator because these were the gold standard items identified from the human produced translations. By using this value, we obtained a coarser response rate (compared to GT) that takes into account both the subject and machine performance. Each of these RT items is treated as being independently identified by each gold standard rater viewing the document.

Chapter 4

Data Analysis Phase 1: Relationship Establishment, Correlation Analysis and Results

Translation evaluation is primarily conducted using one of the various autometric approaches described in chapter 2. These methods approximate translation quality and enable developers to rapidly test changes in their system output. Autometrics have been justified as an evaluation tool based on how well they correlate across an entire document testbed with human judgments of translation quality on the same data set. With system ranking as the primary objective for such translation evaluations, real system users have remained out of the loop.

This dissertation demonstrates the applicability of pre-existing automated evaluation metrics—BLEU, GTM, METEOR, and (o)TER—in determining correct response rate on our information extraction task. Responses collected from the study described in Chapter 3 are analyzed in two phases. The first phase of data analysis involves descriptive summaries, correlation analysis, and permutational testing. Descriptive statistics obtained from scatterplots and nonparametric regression provide us with relationship summaries that enable our further analyses. Correlation

analysis indicates the differences between the strength of relationships among the variables under consideration at corpus versus document level. This analysis motivates the use of a more sophisticated (partial) correlations analysis later on to assess the strength of hit rate by metric relationships within MT $\times$ WH groups of documents. Permutational testing provides a nonparametric method for determining significance of our chosen statistics. The overall purpose of this phase of analysis is to motivate the application of statistical modeling in MT evaluation by exposing the limitations of descriptive statistics, mainly correlation analysis, generally used in this area.

This chapter addresses the following Phase 1 research questions:

1. What is the nature of the relationships between extraction task performance and automated translation evaluation metrics? Is the relationship linear? What are the magnitude and direction of the correlational patterns? Is the relationship statistically significant?

2. Is there a statistically significant difference in the utility/quality relationship of different autometrics (i.e., do certain metrics have evidently stronger relationships with task performance)? Do recodes that average away MT dependence of the metrics provide more defined, less noisy relationships as opposed to task performance than the raw metrics?

This analysis shows the need to extend beyond correlations to utilize autometrics appropriately. That is, the lack of a clear linear relationship at the document level which is the natural level for task handling purposes is demonstrated. In Chap-

ter 6, the second phase of data analysis examines results from the first phase and extends the work by applying generalized linear modeling techniques to the experiment data. The **R** statistical language package [53, 66] was used for all statistical analyses in this thesis.

4.1 Data Description

With 59 subjects each analyzing 18 translated documents, 1062 response cases were collected. However, with one server-client connection crash during the experiment, two instances of translated documents viewed and marked by subjects could not be processed. Additionally, it is shown later in this chapter that one MT-doc combination was found to be an outlier. Taking all this into account, the final data analyses are computed on 1040 cases.

The data were entered into an **R** data frame with each row representing one subject's analysis of one of the translated documents. Table 4.1 shows a portion of the data set for twenty of the experimental cases. The first column denotes the subject who viewed the particular document. The second through fourth columns represent the document identifier detailing the machine system which produced the translation, the WH-type category, and the replicate number (1-6) of the category, respectively. The fifth column is the total number of possible items to extract from the document (RTMTot). The sixth column is the proportion of correct items the subject extracted (Hits) out of the total possible items. The remaining columns correspond to the various autometric scores computed for the document.

Table 4.1: Random sample of 20 cases from data collected. Column headings are described in the text.

Subj	MT	WH	Rep	RTMTot	Hit Rate	BLEU	METEOR	oTER	GTM
S43	3	WHERE	4	7	.5714	.040	.421	.214	.432
S30	1	WHERE	3	6	.8333	.074	.333	.247	.567
S29	1	WHERE	6	8	.6250	.084	.445	.302	.601
S56	3	WHERE	6	8	.5000	.050	.354	.167	.476
S9	3	WHO	6	10	.4000	.126	.582	.333	.584
S3	2	WHO	1	7	.1429	.224	.553	.414	.637
S44	1	WHO	2	7	.4286	.095	.505	.245	.556
S57	2	WHEN	4	12	.0000	.099	.383	.297	.540
S59	1	WHERE	6	8	.3750	.084	.445	.302	.601
S1	2	WHO	3	9	.2222	.211	.503	.423	.611
S46	1	WHEN	5	5	.4000	.043	.198	.220	.401
S55	2	WHERE	6	8	.5000	.155	.494	.355	.669
S34	3	WHO	2	7	.2857	.046	.223	.245	.373
S46	1	WHO	1	7	.2857	.060	.468	.257	.532
S15	2	WHERE	4	7	.5714	.188	.595	.396	.547
S52	2	WHERE	6	8	.8750	.155	.494	.355	.669
S14	2	WHEN	3	8	.3750	.283	.567	.376	.692
S48	3	WHO	3	9	.5556	.024	.407	.227	.446
S28	2	WHERE	2	9	.5556	.139	.523	.332	.501
S56	1	WHERE	4	7	.2857	.020	.227	.106	.392

4.1.1 Summary of Project Data

The task-based study identifies the proportion of correct items out of the RT items extracted by subjects from the translated documents—the hit rate—and computed document scores for each autometric. Collection level autometric scores and task response rates are also compute for the roughly 354 documents across each of the three MT systems. These are the values presented in Tables 4.2, 4.3 and 4.4.

Summary statistics were computed to provide a synopsis of the data used in this study. The minimum, maximum, mean, and median of each variable can be found in Table 4.2. METEOR and GTM scores are slightly more dispersed than the other two metrics. GTM and oTER scores classify more documents as above average translations, with 51% (550) and 50% (531) of the responses, respectively, having higher scores than the mean score. The BLEU scores of 35% (373) responses were higher than the mean BLEU score while this held true for 38% of METEOR scores.

Table 4.2: Summary statistics for study variables.

Variable	Min	Max	Mean	Median	Std. Dev
Hit Rate	0	1	.436	.429	.226
BLEU	.016	.283	.106	.080	.075
GTM	.264	.709	.528	.540	.097
METEOR	.198	.621	.425	.413	.100
oTER	.105	.437	.272	.269	.091

Approximately 377 of the document/subject cases yielded 50% or greater hit rates which means that subjects correctly extracted half or more of the WH items in roughly one-third of the cases. The responses tallied by MT can be found in Table

Table 4.3: Hit rates by MT engine, aggregated over all WH-types, subjects and documents.

	MT-1	MT-2	MT-3
# of Correct Extractions (Hits)	1181	1506	1370
Total # of Possible Correct Responses	3091	3066	3086
Hit Rate	.382	.491	.444

4.3. MT-1 yielded significantly lower rates of correct answers from subjects. Table 4.4 shows that this pattern does not hold true for the metrics scores calculated by MT system. MT-1 has the lowest METEOR and oTER scores while MT-3 has the lowest BLEU and GTM scores. Both subject performance and automated metrics indicate MT-2 is the best translating engine in terms of utility and translation quality.

A chi-square test for equality of all three hit rates on 2 degrees of freedom made under the oversimplifying assumption that all of the observations (x_i, y_i) within tabulated cells are independent identically distributed ($\chi^2 = 74.89, p < .001$), followed by pairwise comparisons of hit rates would lead to the strong conclusion that the hit rate of MT-2 is statistically indistinguishable from MT-3 (i.e., the two systems were approximately equal in performance on this metric) and MT-1 has statistically significantly lower hit rates than each of the other two engines. More details can be found in [68], but the p-values here are not strictly meaningful, which is why I undertake the more complicated analyses in later sections.

[1]Recall that the automated metric scores here are not just the averages of the document-level autometric values but a single corpus-level score computed for the entire collection of documents from each system.

Table 4.4: Automated metric scores by MT engine, aggregated over all WH-types, subjects and documents.

Automated Metric[1]	MT-1	MT-2	MT-3
BLEU	.088	.187	.055
GTM	.529	.617	.453
METEOR	.385	.524	.397
oTER	.221	.370	.233

4.2 Correlation Analysis

This dissertation studies the relationship of task performance results and autometrics by exploring various aspects of correlation between the autometrics and subject responses. Simple correlation analysis between subjective human judgments of translation quality and autometric scores has played a vital role in MT evaluation. Researchers have heavily relied on the latest correlation results to determine both a) which system outperforms other systems the most and b) which autometrics best validate this finding. These results are often reported as Pearson correlation coefficients. However, to allow for possible non-linear structure in the data, our alternative choice is also to calculate the more flexible nonparametric Spearman correlation. In fact, [15] has suggested the use of rank transformation procedures because applied statisticians often encounter real-world problems in which the data clearly does not meet normal distributional assumptions often used to motivate the Pearson correlation.

4.2.1 Pearson versus Spearman Rank Correlation

Pearson correlation is the standard measure of correlation and is computed using the formula:

$$r = \frac{S_{XY}}{\sqrt{S_X^2 S_Y^2}} \tag{4.1}$$

where S_{XY} is the sample covariance of X and Y, ${S^2}_X$ and ${S^2}_Y$ are the sample variances of X and Y respectively. Pearson's r measures the linear association between two variables, which can be small even when there is a clear non-linear relationship between the variables.

The Spearman rank correlation is a distribution-free rank statistic that tests the direction and strength of the relationship between two variables [41]. Both sets of data are ranked from the highest to the lowest with the smallest observation having rank 1 and the largest having rank n. Ranks are averaged in the case of ties. Then, the statistic is defined using the formula:

$$\rho = \frac{\sum_{i=1}^{n} \left(R(X_i, \underline{X}) - \frac{n+1}{2} \right) \left(R(Y_i, \underline{Y}) - \frac{n+1}{2} \right)}{\sqrt{\sum_{i=1}^{n} \left(R(X_i, \underline{X}) - \frac{n+1}{2} \right)^2 \sum_{i=1}^{n} \left(R(Y_i, \underline{Y}) - \frac{n+1}{2} \right)^2}} \tag{4.2}$$

where $R(A_i, \underline{A})$ represents the rank of A_i in the subset $\underline{A}$ and is equal to the number of elements of $\underline{A}$ less than or equal to A_i. This method, based on ranking the two variables, makes fewer assumptions about the distribution of the values and measures monotone rather than only linear covariation.

4.2.2 Correlation in Aggregated Evaluation Datasets

In general, autometrics were computed on a document collection translated by a system, and the collection is then given a system score for each autometric. Separately, humans are solicited to make quality judgments on the documents according to some preassigned numeric scale. Note that human judgments are made at the individual document level and then averaged across the collection to produce a 'system specific' human judgment score, as well. Thus for any comparison, the number of total possible data points in a test set is $S \times (M + 1)$ where S is the number of systems and $M + 1$ is the number of metrics plus the human judgment of the system. For example, if there are 3 systems under consideration (*S1*, *S2*, and *S3*) and 2 different autometrics (*M1* and *M2*) to compare to human judgment scores (*HJ*), there are only 9 total data points. Moreover, the pairwise autometric versus judgment correlation is computed from only 3 data points and in any case, using this aggregated approach, one would only have as many data points to correlate as systems under study.

Consider the system level correlation in Table 4.5 between the autometrics studied in this work and task responses.[2] For most metrics, our results show high correlations in the 'aggregate' sense between autometrics and task performance similar to those observed between autometrics and human judgments [52, 62, 39, **?**]. At this level of analysis, the METEOR and oTER autometrics correlate highly with human performance on this task. However, as is true of previous work, the calculated

[2]For the sake of comparison with previous results, only Pearson correlation is used in this section for correlating aggregate scores.

Table 4.5: Autometrics correlated with hit rate for aggregate scores by MT using Pearson correlation

BLEU	GTM	METEOR	oTER
.6634	.4676	.8654	.8626

'correlation' simply represents the normalized inner product of only 3-dimensional vectors (3 metric values, one for each MT system) for the whole collection.

The conclusions generated from aggregate-level correlations are not useful for further interpretation because they are a very coarse summary of group differences. However, more definitive conclusions about the relationship are possible when more system-level data points are available. Despite this caution, the most widely used autometric—BLEU—gained its popularity mainly because it has been shown to correlate highly with human judgments at the aggregated system level [52]. Moreover, the rankings of Machine Translation (MT) systems produced by both BLEU scores and human judgments were shown to be the same. BLEU's creators, Papineni et al. [52] assert that the five MT system BLEU scores obtained from a 500 translated sentence collection in their study have correlations greater than .96 with both monolingual and bilingual sets of human judgments on the same 500-sentence collection from the same five systems.

While aggregating scores across collections has given system developers insight into the development of their systems on average for particular collections of documents, users have not been able to make the connections between individual document scores and autometrics. Some attempts have been made to utilize autometrics at other levels [50, 7, 58]. As one might expect, lower levels of aggregation

have not achieved the high correlations that were observed in the aggregated case with BLEU, but recently some metrics have done well in comparison with others. For instance, METEOR[39] and CDER [42] were purposely designed to improve correlation at the *sentence* level, and the authors of both have shown results of higher correlations at the sentence level with human judgments than other metrics.

Since our goal is to find a relationship between autometrics and subject task performance, and eventually to calibrate **document scores** with some degree of utility for a specific task, it is relevant that the data is analyzed at a level more useful for task-based comparisons. Next, the differences in correlation results between task responses and autometrics in the study at the individual document level of aggregation are shown.

4.2.3 Correlation at Unit Level for Task Performance Evaluation

Although correlation at the aggregated (system) level may be sufficient for quick system evaluation, when trying to make a strong connection between the utility of the translated documents and autometrics, I want to answer a question that has not been addressed before: *what happens to the relationship between autometrics and response rates when scores are compared at the document level?* To test this question, the entire set of *non-aggregated* 1060 individual document scores is compared. It is found that, even though the system rankings are the same, the degree of correlation between each autometric and task performance drastically changes [see Table 4.6] from the aggregate-level results. This suggests that useful relation-

ships between autometrics and task performance at document level may not exist. However, scatterplot smoothing in the next section indicates quite the contrary.

Table 4.6: Autometric correlation with hit rate on non-aggregate individual document scores for the 1060-document set

Method	BLEU	GTM	METEOR	oTER
Pearson r	.153	.209	.232	.223
Spearman ρ	.211	.193	.242	.231

Reeder and White [54] mention that for many reasons, the evaluation issue is not solved since finer-grained metrics for smaller units of data (i.e., sentences, documents, etc) are needed. This is true especially in this dissertation because I try to use the metrics to **predict** task performance at the document level. The remainder of this dissertation shows that although a weak correlation exists at this stage of granularity, that does not necessarily indicate that there is no relationship between the two variables. One metric may prove to be a better *predictor* of task performance although its correlation may be weak. I proceed by using data smoothing techniques as well as the patterns for correlation, scatter, and linearity in the relationship of variables cross-classified into finer groups to identify which of the given set of metrics may be better in predicting our extraction task responses.

4.3 Data Smoothing Techniques

I begin to analyze each of the autometrics' relationship with task performance in this experiment by asking the most model-free questions. Can anything be gained from this document level data without the use of rigorous statistical models? This

will be determined by how much noise there is in the data.

Suppose $(x_i, y_i), i = 1, ..., n$ are observations with

$$Y_i = f(x_i) + \epsilon_i.$$

The nonparametric regression problem, as described in [28], is considered to be an estimate of the structure of the unknown smooth function f, where the ϵ_i are independent errors with mean 0. If the observations

$$(x_1, y_1), ..., (x_n, y_n)$$

$$x_1 < x_2 < ... < x_n$$

are points on a scatterplot, the goal of a scatterplot smoother would be to fit a smooth function that describes the dependence y has on x. The smoothed points are $(x_i, \hat{y}_i)$ where $\hat{y}_i$, the fitted value at x_i, estimates the location of Y given $X = x_i$. A nonparametric regression of Yon X is formed by plotting the smoothed points, ultimately making the tendency of the cloud of points on the scatterplot become more apparent.

There are several nonparametric regression techniques for smoothing data including: local averaging, kernel estimation, and smoothing splines. I focus on a widely used method for smoothing scatterplots of noisy data, originally introduced by [13], called *locally weighted scatterplot smoothing*. This method builds on classical approaches, such as linear and nonlinear least squares regression, as well as kernel estimation, providing weighted combinations of simple models fitted to localized subsets of the data.

This very useful technique, referred to as "lowess" or "loess," partitions the data into successive windows of x values, to fit least-squares regression lines to the x versus y points within those windows, and to find a smooth curve fitting those pieces of fitted lines together across the whole x axis. This gives an overall display of curvilinear average dependence in settings where no overall linear relationship and possibly not even a monotonic relationship between x and y variables exists.

The general steps for obtaining a smoothed value for a given x_0 are as follows [24]:

1. Choose the *span* (fraction of data around x_0), $0 < s < 1$, to include in each fit.[3]

2. Calculate regression weights for each value of x_i in the above span using the formula.[4]

$$w_i = W\left[\frac{x_i - x_0}{h_i}\right].$$

3. Fit the local regression equation

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 {x_i}^2 + ... + \beta_k {x_i}^k + \epsilon_i$$

to minimize:

$$\sum w_i {\epsilon_i}^2$$

[3]The larger the s value the smoother the fit. The default value for s in the **R** version of this algorithm is $\frac{2}{3}$.

[4]"W" here is referred to as the tricube weight function,

$$W(z) = \begin{cases} (1 - |z|^3)^3 & \text{for } |z| < 1 \\ 0 & \text{for } |z| \leq 1 \end{cases}$$

which gives greater weight to x_i closest to x, and h_i is the half-width of the points surrounding x_i.

using weighted least squares regression. Let $\hat{\beta}_j(x_0)$ denote the estimated regression coefficients.

4. Compute fitted value using the equation:

$$\hat{Y}_0 = \beta_0(x_0) + \hat{\beta}_1(x_0)x_0 + ... + \hat{\beta}_k(x_0){x_0}^k$$

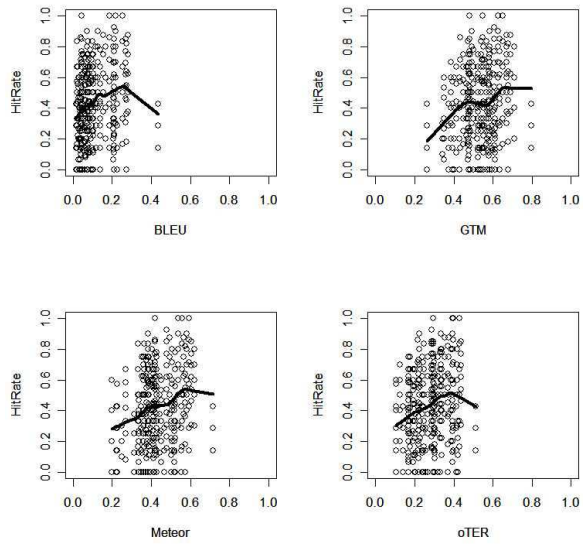
Finally, local regressions are estimated using this same procedure for other x-values, and the fitted values are connected in a nonparametric regression curve.

4.3.1 Relationship Summaries

Scatterplots of each metric plotted against the proportion of correct answers (hit rate) are displayed in Figure 4.1. Initially, the noisy raw data indicate that there is no clear picture of a possible association between metrics and task responses. The lowess smoothing technique described in the previous section enhances the interpretation of the plot. The bold line shows that in each case, there is a generally increasing pattern of scores with an increase in hit rate.

It is interesting to point out in Figure 4.1 that there are three outlying points in the data across all autometrics. These three points represent the same WHEN document from MT-2 that was viewed by a total of 20 subjects. Each of the autometrics achieve extremely higher scores for this document versus other documents in the collection while its hit rates are slightly lower than the average hit rates. This particular document is considered an anomaly since the scores are so extreme compared to other documents and there is no evident reason for this peculiarity. Entries of it are omitted from further study and the remainder of our analysis will

Figure 4.1: Scatterplot of the relationship between autometric scores and hit rate with smoothed lines denoting the lowess scatterplot smoother



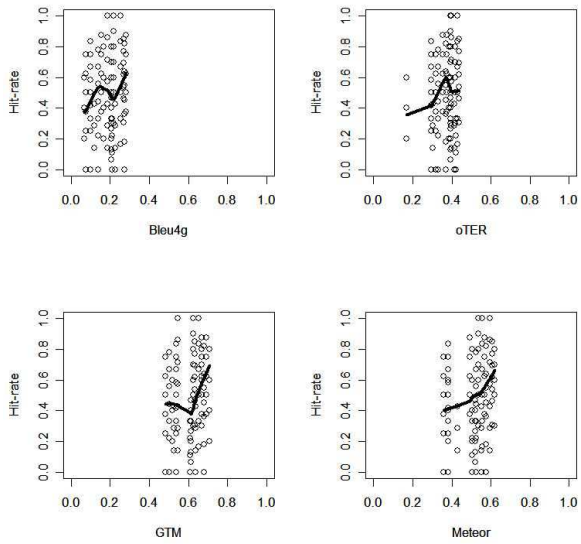
be done on 1040 cases rather than the original 1060 cases to prevent results from being adversely affected by this phenomenon. Table 4.7 shows that the correlation between metrics and task performance from Table 4.6 slightly increases once this outlier is removed. This increase is more apparent for the correlations involving BLEU and GTM that were much lower than the others. Now all metrics appear to be on equal footing in relation to hit rate.

Table 4.7: Autometric correlation with Hit rate on individual document scores with outlier document removed.

Method	BLEU	GTM	METEOR	oTER
Pearson r	.252	.269	.297	.281
Spearman ρ	.249	.230	.281	.270

While scatterplot smoothers are a good tool for lots of data, they are not as good for smaller sets cross-classified into finer categories. For instance, there is less

Figure 4.2: Scatterplot of automated metric scores versus Hit rate with lowess lines for MT system 2



scatter in the relationship cross-classified by MT system 2, but Figure 4.2 shows that the lowess curves are more nonlinear and are nonmonotonic. When the data is observed by WH-type and further cross-classified into MT $\times$ WH, similar patterns can be found. Overall, lowess lines for the plots with BLEU and GTM metrics versus hit rate are nonlinear but for METEOR and oTER, there is a mostly linear pattern. The latter two metrics have more readily visible increasing tendencies with hit rate than the former two.

The noisy and mixed patterns of the scatterplots when the data are cross-classified by MT-2 indicate that the relationship seen in Figure 4.1 may be due to the MT variable effect on the automated metric scores. This finding is more confirmatory than surprising because previous evaluation work has shown that autometrics are useful in distinguishing between MT systems of varying quality. It was

showed in the summary of experimental results in Section 4.1.1 that MT systems can also be distinguished based on utility via the task response rates. Yet it is not evident whether autometrics contribute anything beyond being able to distinguish between systems when it comes to task response rates. Further analysis will help refine these distinctions.

4.4 Further Correlation Analysis

It has been established in this work that there is a positive and generally monotonic relationship between autometric and task performance variables in our data. However, the evidence of a strong relationship in the presence of other effects—such as MT and WH-type—is less apparent. This leads to further correlation analysis, in which I extend beyond the study of the strict bivariate relationships by using other categorical variables in the cross-classified data set to determine the extent of residual correlation between the two variables once the third variable is held constant.

In this section, I want to know: *to what extent does the autometric score still account for the correct task response rate after adjusting for the MT effect and is this different for different MT systems?* I also go further to explore: *are there any groups other than MT that show interesting document to document variation that will help quantify the response rate?* Studying the within-group and partial relationship answers such questions better than does population-wide correlation. Permutation tests, discussed in the next section, are used to determine whether relationships are

significant.

4.4.1 Permutation Tests

Permutation tests provide a robust nonparametric alternative to using traditional, model dependent significance testing methods. The main idea of permutational testing [26] is to estimate the empirical distribution of statistic values over the ensemble of randomly permuted datasets. Sampling of the permuted data provides a null hypothesis benchmark and "exact" significance level. Permutation tests of significance are conducted in this dissertation according to the following procedure [27]:

1. Compute the correlation of the original observations.

2. Resample the autometric scores, based only on permutations that preserve groups, and recompute the correlation for these permuted values.

3. Calculate the exact significance level of the test from

$$\text{p-value} = \frac{(\#\text{recomputed statistics} \geq \text{original statistic})}{n}$$

   where $n = 5000$ is the number of permutations.

In this research, permutational procedures not only facilitate computing the p-value for testing whether the correlations are significantly greater than zero in the autometrics versus hit rate relationship, but also they aid in summarizing adequate models achieved in the next phase.

4.4.2 Within-Group Correlation

The results of the within MT-group correlation between hit rate and autometrics can be found in Table 4.8. Documents translated by all MT systems showed monotonic and significantly positive associations between hit rate and evaluation metric scores across all autometrics. Thus, there are real within MT-group relationships in our data. Response rates for extraction within WH-group also showed a monotonic and statistically significant positive association with evaluation metric scores across all metrics. Thus similarly, Table 4.9 shows that there are significant relationships after grouping the data by WH-type.

Table 4.8: Autometric correlation with hit rate on individual document scores cross-classified by MT system. Permutational significance values are shown in parentheses.

Method	MT	BLEU	GTM	METEOR	oTER
Pearson r	1	.156(.002)	.161(.003)	.115(.03)	.203(<.001)
	2	.169(.01)	.256(<.001)	.273(<.001)	.151(.04)
	3	.306(<.001)	.358(<.001)	.306(<.001)	.234(<.001)
Spearman ρ	1	.140(.009)	.147(.005)	.109(.04)	.235(<.001)
	2	.134(.002)	.303(<.001)	.298(<.001)	.111(.006)
	3	.298(<.001)	.323(<.001)	.311(<.001)	.182(<.001)

When documents are further classified into the 3×3 (WH $\times$ MT) grouping, it appears in several of the cases in Table 4.10, with the exception of WHO documents, the relationship between metrics and hit rate is negative thus, inconsistent.[5]Also, more relationships are found to be non-significant at this finer classification. Yet in some cases, there are very strong relationships between hit rate and autometric as demonstrated by the Spearman correlation value of GTM (.754) for WHO

[5]Recall that there are about 115 data points within each MT $\times$ WH group in this table.

Table 4.9: Autometric correlation with hit rate on individual document scores cross-classified by WH type. Permutational significance values are shown in parentheses.

Method	WH	BLEU	GTM	METEOR	oTER
Pearson r	WHO	.246(<.001)	.342(<.001)	.347(<.001)	.235(<.001)
	WHERE	.220(<.001)	.221(<.001)	.273(<.001)	.268(<.001)
	WHEN	.316(<.001)	.244(<.001)	.232(<.001)	.372(<.001)
Spearman ρ	WHO	.253(<.001)	.292(<.001)	.320(<.001)	.198(<.001)
	WHERE	.229(<.001)	.185(.001)	.252(<.001)	.251(<.001)
	WHEN	.250(<.001)	.158(.002)	.237(<.001))	.331(<.001)

documents from MT-3.

This section showed that autometrics reflect task performance rates even within different cross-classifications of our data. However, this relationship is not always consistent. In the next section, partial correlations reveal the population-wide association after removing the MT $\times$ WH group effects.

4.4.3 Partial Correlation

If X_1, X_2, and X_3 are three random variables, the partial correlation coefficient of the variables can be calculated by the formula

$$r_{x_1x_2.x_3} = \frac{r_{x_1x_2} - r_{x_2x_3}r_{x_1x_3}}{\sqrt{(1 - {r_{x_2x_3}}^2)(1 - {r_{x_1x_3}}^2)}} \tag{4.3}$$

where $r_{x_1x_2}$, $r_{x_2x_3}$, and $r_{x_1x_3}$ are the ordinary Pearson r correlation coefficients obtained between the indicated pairs of variables [14]. I am interested in the *Spearman* version of this partial correlation as introduced by [34], but expanded to the case were variable x_3 is a discrete grouping effect. This correction after adjusting for a grouping effect does not seem to have been studied in the literature.

Such partial correlations are sought for the categorical variable Z representing

Table 4.10: Autometric correlation with Hit rate on non-aggregate individual document scores cross-classified by WH × MT type. Permutational significance values for non-significance at the .05 level are shown in parentheses.

Method	WH	MT	BLEU	GTM	METEOR	oTER
Pearson	WHO	1	.378	.353	.248	.202
		2	.350	.389	.524	.240
		3	.493	.698	.519	.541
	WHERE	1	.189	.256	.160(.07)	.284
		2	-.174(.07)	.075(.42)	.014(.88)	-.017(.86)
		3	.028(.76)	.138	-.172(.07)	-.182
	WHEN	1	-.299	-.205	-.107(.24)	.053(.57)
		2	.461	.454	.451	.416
		3	.398	.273	.263	.223
Spearman	WHO	1	.571	.322	.217	.123(.18)
		2	.161(.09)	.285	.467	.172(.07)
		3	.472	.754	.694	.478
	WHERE	1	.195	.274	.116(.20)	.322
		2	-.227	.094(.29)	-.020(.83)	-.012(.89)
		3	.006(.95)	.030(.75)	-.128(.17)	-.111(.24)
	WHEN	1	-.220	-.161(.08)	-.098(.28)	.074(.41)
		2	.528	.456	.502	.366
		3	.492	.311	.184	.196

the MT, WH, and MT × WH cross-classified groups. Partial correlations can be thought of as a way to reveal the population-wide correlation after removing the MT × WH group effects. Our use of partial correlation in this work offers a more general methodological perspective of the partial rank correlation statistic since when considering Z, there is a decision that can be made as to how to actually rank the data. In general, this within Z stratum detection of relevant variable relationships is appropriate for many applications such as ecological and social science studies.

I derive two distinct expressions yielding two different statistics for partial rank correlation after adjusting for categorical grouping effects.

Method 1: ρ computed from X and Y, linearly corrected for Z by the group mean.

This statistic is called S_1. Let $I_Z(z)$ denote the indicator function such that

$$|x| = \begin{cases} 1 & \text{if } z \in Z; \\ 0 & \text{if } z \notin Z. \end{cases}$$

then $X_i^* = X_i - c_{Z_i}, Y_i^* = Y_i - d_{Z_i}$ and

$$S_1 = \frac{Cov\left[R(X_i{}^*, \underline{X}^*)R(Y_i{}^*, \underline{Y}^*)\right]}{\sqrt{Var(R(X_i{}^*, \underline{X}^*))Var(R(Y_i{}^*, \underline{Y}^*))}} \tag{4.4}$$

for $j = 1, ..., L$ indexing the different levels of Z, $c_j = \sum_{i=1}^n I_{[Z_i=j]}X_i / \sum_{i=1}^n I_{[Z_i=j]}$ and similarly $d_j = \sum_{i=1}^n I_{[Z_i=j]}Y_i / \sum_{i=1}^n I_{[Z_i=j]}$. Recall that $R(X_i{}^*, \underline{X}^*)$ and $R(Y_i{}^*, \underline{Y}^*)$ are the ranks of the corresponding X and Y values. This formula can be equivalently written as

$$S_1 = \frac{\sum_{j=1}^n \left(R(X_i{}^*, \underline{X}^*) - \frac{n+1}{2}\right)\left(R(Y_i{}^*, \underline{Y}^*) - \frac{n+1}{2}\right)}{\sqrt{\sum_{j=1}^n \left(R(X_i{}^*, \underline{X}^*) - \frac{n+1}{2}\right)^2 \sum_{j=1}^n \left(R(Y_i{}^*, \underline{Y}^*) - \frac{n+1}{2}\right)^2}} \tag{4.5}$$

Method 2: Compute weighted combination of within Z-group ρ. This statistic is called S_2. Let w_z be the weight given for group z and $\underline{X}^{(z)} = (x_i : i \in J_z)$ where J_z is the subset of indices i = 1, ...,n for which $Z_i = z$. Then,

$$S_2 = \frac{\sum_{z=1}^L w_z \rho_{y,x|Z=z}}{\sum_{z=1}^L w_z} \tag{4.6}$$

where ρ in this case is defined as:

$$\rho_{y,x|Z=z} = \frac{\sum_{i \in J_z} \left(R(X_i, \underline{X}^{(z)}) - \frac{n+1}{2}\right)\left(R(Y_i, \underline{Y}^{(z)}) - \frac{n+1}{2}\right)}{\sqrt{\sum_{i \in J_z} \left(R(X_i, \underline{X}^{(z)}) - \frac{n+1}{2}\right)^2 \sum_{i \in J_z} \left(R(Y_i, \underline{Y}^{(z)}) - \frac{n+1}{2}\right)^2}} \tag{4.7}$$

Table 4.11: Autometric partial rank correlation with hit rate on non-aggregate individual document scores by grouping effect. All values are permutational significant with p-value equal to .001.

Grouping Factor	Statistic	BLEU	GTM	METEOR	oTER
MT	S_1	.156	.251	.232	.178
	S_2	.193	.255	.237	.179
WH	S_1	.243	.241	.273	.270
	S_2	.244	.214	.271	.257
WH $\times$ **MT**	S_1	.200	.300	.243	.183
	S_2	.200	.256	.204	.174

4.4.4 Permutational Significance of Partial Correlation for Task Performance Evaluation

The results of the partial correlation between hit rate and autometrics in the presence of the MT, WH, WH $\times$ MT grouping effects, respectively, for statistics S_1 and S_2 can be found in Table 4.11. The partial correlations between task performance and autometrics after accounting for groups are slight but all partial relationships are significant permutationally with p-values equal to .001.

Chapter 5 further expands the idea of partial Spearman rank correlation and permutation testing by looking at a simulation study to investigate the characteristics of statistics S_1 and S_2, above, when permutation tests are conducted for the conditional bivariate-normal model given Z.

4.5 Recoding Predictor Variables

An additional approach to determine whether autometrics can play a role in task based evaluation is to recode the metrics to remove their dependence on the MT

system and to use the new variable to test if the metrics have additional information concerning the relationship with task response rate. The method proposed is to take each metric and find a non-MT dependent variant. This is achieved by averaging across MT systems the ratios of autometric scores divided by the within-MT average over documents.

4.5.1 Metric Average Variable Recode

The experimental data is classified into 3 groups based on MT and calculate autometric *average* scores: BLEUavg, GTMavg, METEORavg, and oTERavg in the following manner. Let $i = 1, .., 18$ index document, let $j = 1, 2, 3$ index MT system, and let the document scores for a particular metric be denoted u_{ij}. Then the system average scores, A_j, are given by $A_j = \bar{u}_{\cdot j} = (1/18) \sum_{i=1}^{18} u_{ij}$ after which the adjusted document scores are $AD_{ij} = u_{ij}/A_j$, and the final recoded scores become

$$u_{avg,i} = (1/3) \sum_{j=1}^{3} AD_{ij}. \tag{4.8}$$

4.5.2 Example

An example of the autometric average calculation is shown for BLEU using the 18 original document scores [see Table 4.12]. S_i is computed in step one of our method to obtain the MT system average scores found on the last row of the table. These values are used to calculate the new document score adjusted by its MT system average score found in Table 4.13. Finally, the recoded BLEUavg variable in column six is computed by dividing column 5 (the sum of the adjusted scores across

MT systems) by the number of systems. Following this procedure, a document specific recode of all the autometrics that has been adjusted for all instances of the translated document across all MT systems results. The recoding of metrics will be used in the next phase to compare against initial variables in the modeling.

Table 4.12: Document BLEU scores for each MT system

Document	MT1	MT2	MT3
When-1	.137	—	.078
When-2	.103	.074	.038
When-3	.105	.283	.073
When-4	.063	.099	.045
When-5	.043	.069	.034
When-6	.091	.165	.115
Where-1	063	.272	.06
Where-2	.104	.139	.08
Where-3	.074	.255	.021
Where-4	.02	.188	.04
Where-5	.038	.208	.027
Where-6	.084	.155	.05
Who-1	.06	.224	.016
Who-2	.095	.119	.046
Who-3	.087	.211	.024
Who-4	.209	.27	.044
Who-5	.071	.216	.043
Who-6	.039	.206	.126
System Average(S_i)	.083	.186	.053

4.6 Phase 1 Summary

This chapter has examined the relationships between WH-extraction task response rates and autometrics by utilizing several descriptive statistics including an in-depth correlation analysis. Most of the statistical tools discussed in this chapter are not original except for the partial rank correlation statistics. However, my

Table 4.13: MT system adjusted BLEU scores, document summed score, and final recoded BLEUavg score

Document	AD_1	AD_2	AD_3	$\sum_{j=1}^{3} AD_j$	Final BLEUavg
When-1	1.6595	—	1.4625	3.1220	1.5610
When-2	1.2476	0.3990	0.7125	2.3591	0.7864
When-3	1.2719	1.5258	1.3688	4.1665	1.3888
When-4	0.7631	0.5338	0.8438	2.1406	0.7135
When-5	0.5209	0.3720	0.6375	1.5304	0.5101
When-6	1.1023	0.8896	2.1563	4.1482	1.3827
Where-1	0.7631	1.4665	1.1250	3.3547	1.1182
Where-2	1.2598	0.7494	1.5000	3.5092	1.1697
Where-3	0.8964	1.3749	0.3938	2.6650	0.8883
Where-4	0.2423	1.0136	0.7500	2.0059	0.6686
Where-5	0.4603	1.1215	0.5063	2.0880	0.6960
Where-6	1.0175	0.8357	0.9375	2.7907	0.9302
Who-1	0.7268	1.2077	0.3000	2.2345	0.7448
Who-2	1.1507	0.6416	0.8625	2.6549	0.8850
Who-3	1.0538	1.1376	0.4500	2.6415	0.8805
Who-4	2.5316	1.4558	0.8250	4.8124	1.6041
Who-5	0.8600	1.1646	0.8063	2.8309	0.9436
Who-6	0.4724	1.1107	2.3625	3.9456	1.3152

application of correlational and permutational tools here, and in modeling-building presented in Chapter 6, were customized for this particular MT evaluation application and serve as a case study without any precursor in the MT literature.

It was shown that autometric sensitivity to granularity can be exposed when trying to assess task performance. This study calls for document-level autometrics. It is found that even though correlations are quite low at this level, there is a slight relationship when autometrics are considered within the cross-classifications of other variables in the study, namely the MT system that translated a particular document or the actual WH-task at hand. There is certainly variety in these relationships, group by group, and even though they are hard to see by eye and may be weak,

permutational testing shows that the relationship is one that cannot be ascribed to chance. These findings motivate the logistic regression models in Phase 2 by informing us of predictive variables to test. The use of grouping variables in this study suggests that partial Spearman rank correlation can be used in assessing the data in Phase 2. The next chapter discusses the asymptotic behavior and power of our Spearman rank correlation statistics.

Chapter 5

A Look at Partial Rank Correlation within Groups

Let X, Y and Z be random variables for which X and Y are jointly distributed given Z, where Z is a discrete group indicator. How can the true degree of relationship between X and Y conditionally given Z be summarized? A good general class to consider is

$$f_{X,Y|Z}(x,y|z) = f_0\left(\frac{x-\mu_z}{\sigma_z}\right) \cdot g_0\left(\frac{y-\mu_y-\rho_z\frac{\sigma_y}{\sigma_z}(x-\mu_z)}{\sigma_y\sqrt{1-{\rho_z}^2}}\right)$$

In this chapter, two different expressions are derived for partial rank correlation: (1) S_1—a rank correlation computed from X and Y after linearly correcting by their group means within levels of Z, and (2) S_2—a weighted combination of within Z-group rank correlations.

The question to investigate is how the statistics (S_1 and S_2) in equations (4.5) and (4.7), respectively, perform asymptotically in detecting within group relationships. Thus, the asymptotic properties are approximated under bivariate normality. The power of each statistic to detect real within group relationships for a broad

range of within group dependence models is determined. Theoretical expectations of these statistics are presented and simulations of their relative performance are shown for various patterns of conditional correlations as a function of Z. Percentage points of the permutational distribution are estimated.

5.1 Method 1: Rank Correlation Linearly Corrected by Group Mean

As stated in section 4.4, the statistic S_1 can be written as

$$S_1 = \frac{\sum_{i=1}^{n} (R({X_i}^*, \underline{X}^*) - \overline{R({X_i}^*, \underline{X}^*)})(R({Y_i}^*, \underline{Y}^*) - \overline{R({Y_i}^*, \underline{Y}^*)})}{\sqrt{\sum_{i=1}^{n} (R({X_i}^*, \underline{X}^*) - \overline{R({X_i}^*, \underline{X}^*)})^2 \sum_{i=1}^{n} (R({Y_i}^*, \underline{Y}^*) - \overline{R({Y_i}^*, \underline{Y}^*)})^2}} \tag{5.1}$$

Recall that, $I_Z(z)$ denotes the indicator function, $X_i^* = X_i - c_{Z_i}, Y_i^* = Y_i - d_{Z_i}$ and for j=1,...,L indexing the different levels of Z, $c_j = \sum_{i=1}^{n} I_{[Z_i=j]} X_i / \sum_{i=1}^{n} I_{[Z_i=j]}$ and $d_j = \sum_{i=1}^{n} I_{[Z_i=j]} Y_i / \sum_{i=1}^{n} I_{[Z_i=j]}$. The values of c_z and d_z that are used to correct group $Z = z$ are not known in advance as constants and must be supplied as statistical estimators.

Focusing on the numerator of equation (5.1), the following derivation obtains.

$$\sum_{i=1}^{n}(R(X_i^*,\underline{X}^*)-\overline{R(X_i^*,\underline{X}^*)})(R(Y_i^*,\underline{Y}^*)-\overline{R(Y_i^*,\underline{Y}^*)})$$

$$\begin{aligned}
&= \sum_{i=1}^{n}(R(X_i^*,\underline{X}^*)-\frac{n+1}{2})(R(Y_i^*,\underline{Y}^*)-\frac{n+1}{2})\\
&= \sum_{i=1}^{n}[R(X_i^*,\underline{X}^*)R(Y_i^*,\underline{Y}^*)-\frac{n+1}{2}[R(X_i^*,\underline{X}^*)+R(Y_i^*,\underline{Y}^*)]+\frac{(n+1)^2}{4}]\\
&= \sum_{i=1}^{n}R(X_i^*,\underline{X}^*)R(Y_i^*,\underline{Y}^*)-\frac{n+1}{2}[n(n+1)]+\frac{n(n+1)^2}{4}\\
&= \sum_{i=1}^{n}R(X_i^*,\underline{X}^*)R(Y_i^*,\underline{Y}^*)-\frac{2n(n+1)^2}{4}+\frac{n(n+1)^2}{4}\\
&= \sum_{i=1}^{n}R(X_i^*,\underline{X}^*)R(Y_i^*,\underline{Y}^*)-\frac{n(n+1)^2}{4}
\end{aligned}$$

Similarly the denominator can be simplified as follows:

$$\sum_{i=1}^{n}(R(X_i^*,\underline{X}^*)-\overline{R(X_i^*,\underline{X}^*)})^2$$

$$\begin{aligned}
&= \sum_{i=1}^{n}(R(X_i^*,\underline{X}^*)^2-(n+1)\sum_{i=1}^{n}R(X_i^*,\underline{X}^*)+\frac{n(n+1)^2}{4}\\
&= \frac{n(n+1)(2n+1)}{6}-\frac{n(n+1)^2}{2}+\frac{n(n+1)^2}{4}\\
&= \frac{n(n+1)[2(2n+1)-6(n+1)+3(n+1)]}{12}\\
&= \frac{n(n^2-1)}{12}
\end{aligned}$$

Hence, the formula for S_1 can be rewritten as:

$$S_1=\left(\sum_{i=1}^{n}R(X_i^*,\underline{X}^*)R(Y_i^*,\underline{Y}^*)-\frac{n(n+1)^2}{4}\right)\Big/\frac{n(n^2-1)}{12} \tag{5.2}$$

To find the limiting distribution of S_1, it suffices to know the behavior of $\sum_{i=1}^{n}R(X_i^*,\underline{X}^*)R(Y_i^*,\underline{Y}^*)$. It is known from the Law of Large Numbers that for large samples, the sample mean, $(1/n)\sum_{i=1}^{n}X_i$ of independent identically distributed random variables $X_1, X_2, ..., X_n$ converges to $E(X_i)$, the true mean of the distribution

[11]. Furthermore, $R(X_i^*, \underline{X}^*)/n \approx F_X(X_i^*)$. Thus, it follows that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{R(X_i^*,\underline{X}^*)}{n}\frac{R(Y_i^*,\underline{Y}^*)}{n} \approx E(F_X(X_i^*)F_Y(Y_i^*))$$

and

$$\begin{aligned}
&\sum_{i=1}^{n} R(X_i^*,\underline{X}^*)R(Y_i^*,\underline{Y}^*) \\
&\approx n^3\sum_{i=1}^{n} F_X(X_i^*)F_Y(Y_i^*) \\
&\approx n^3\int\int F_X(x)F_Y(y)f_{X,Y}(x,y)dxdy
\end{aligned} \tag{5.3}$$

Now,

$$\begin{aligned}
F_X(X_i^*) &= P(X_i^* \le x) = P(X - c_z I_Z(z) \le x) \\
&= \sum_{z=1}^{n_z} P(X - c_z I_Z(z) \le x | Z = z)\cdot P(Z = z) \\
&= \sum_{z=1}^{n_z} P(X \le x + c_z | Z = z)\cdot P(Z = z) \\
&= \sum_{z=1}^{n_z} \Phi\left(\frac{x + c_z - \mu_{x,z}}{\sigma_{x,z}}\right)\cdot P(Z = z)
\end{aligned}$$

where $I_Z(z)$ is the indicator variable for group z and n_z is the number of elements in group z. Similarly, $F_Y(Y_i^*) = \sum_w \Phi((y + d_w - \mu_{y,w})/\sigma_{y,w})\cdot P(W = w)$. So that I can make specific calculations and simulations, the rest of this chapter assumes that the distribution of (X, Y) is bivariate normal given Z. From this it follows that,

$$\begin{aligned}
f_{X,Y}(x,y) &= \frac{1}{2\pi\sqrt{1-\rho_z}}\exp\left\{-\frac{1}{2(1-{\rho_z}^2)}\left[\left(\frac{x + c_z - \mu_{x,z}}{\sigma_{x,z}}\right)^2\right.\right. \\
&\quad -2\rho_z\left(\frac{x + c_z - \mu_{x,z}}{\sigma_{x,z}}\right)\left(\frac{y + d_z - \mu_{y,z}}{\sigma_{y,z}}\right) \\
&\quad \left.\left.+\left(\frac{y + d_z - \mu_{y,z}}{\sigma_{y,z}}\right)^2\right]\right\}.
\end{aligned}$$

The right hand side of equation (5.3) above becomes

$$n^3 \int\int \sum_z \sum_v \sum_w P(Z=z)P(V=v)P(W=w)\Phi\left(\frac{x+c_v-\mu_{x,v}}{\sigma_{x,v}}\right)$$
$$\times\Phi\left(\frac{y+d_w-\mu_{y,w}}{\sigma_{y,w}}\right)\frac{1}{2\pi\sqrt{1-\rho_z}}\exp\left\{\frac{-1}{2(1-{\rho_z}^2)}\left[\left(\frac{x+c_z-\mu_{x,z}}{\sigma_{x,z}}\right)^2\right.\right. \tag{5.4}$$
$$\left.\left.-2\rho_z\left(\frac{x+c_z-\mu_{x,z}}{\sigma_{x,z}}\right)\left(\frac{y+d_z-\mu_{y,z}}{\sigma_{y,z}}\right)+\left(\frac{y+d_z-\mu_{y,z}}{\sigma_{y,z}}\right)^2\right]\right\}dxdy$$

A change of variable simplifies this formula. Let

$$\sigma_{x,z}S = X + c_z - \mu_{x,z}$$

$$\sigma_{x,z}T = Y + d_z - \mu_{y,z}.$$

Then

$$\begin{aligned} S,T &\sim N\left(0,\begin{pmatrix} 1 & \rho_z \\ \rho_z & 1 \end{pmatrix}\right) \\ T-\rho_z S &\sim N(0,1-\rho_z^2) \\ T^* &= \frac{T-\rho_z S}{\sqrt{1-\rho_z^2}} \\ (S,T^*) &\sim N(0,1). \end{aligned}$$

With these transformations, equation (5.4) can be written as

$$n^3 \int\int \sum_{z,v,w} P(Z=z)P(V=v)P(W=w)\Phi\left(\frac{\mu_{x,z}-c_z+c_v-\mu_{x,v}+\sigma_{x,z}s}{\sigma_{x,v}}\right)$$
$$\Phi\left(\frac{\mu_{y,z}-d_z+d_w-\mu_{y,w}+\sigma_{y,z}(\rho_z s+\sqrt{1-\rho_z^2}t^*)}{\sigma_{y,w}}\right)\frac{1}{2\pi\sqrt{1-\rho_z}}$$
$$\exp\left(\frac{-s^2+(\rho_z s+\sqrt{1-\rho_z^2}t^*)^2-2\rho_z s(\rho_z s+\sqrt{1-\rho_z^2}t^*)}{2(1-{\rho_z}^2)}\right)$$

Note that $c_z = \mu_{x,z}$ and $d_z = \mu_{y,z}$ because the former are large sample consistent estimators of the latter. Finally, this mixture of normal random variables can be written in the form,

$$n^3 \int\int g(s,t^*)\phi(s)\phi(t^*)dsdt^* \tag{5.5}$$

where $g(s,t^*) = \sum_z \sum_v \sum_w P(Z=z)P(V=v)P(W=w)\,\Phi(\sigma_{x,z}s/\sigma_{x,v})\,\Phi((\sigma_{y,z}/\sigma_{y,w})$ $(s\rho_z + \sqrt{1-\rho_z^2}\,t^*))$. This can be approximated using a well-known *Gaussian quadrature* formula

$$\sum_{ij} g(p_i\sqrt{2}, p_j\sqrt{2})\frac{w_i w_j}{\pi}. \tag{5.6}$$

where p_i and w_i are the computed nodes and weights respectively. Once the number of nodes is chosen, the locations and weights are uniquely determined and provide well-controlled high accuracy for the integrals in equation (5.5). This approach to numerical integration will work whenever the function g being integrated can be approximated accurately by a polynomial over the range of integration [57].

5.2 Method 2: Weighted Sum of Rank Correlations within Groups

Statistic S_2 is defined as

$$S_2 = \frac{\sum_{z=1}^{L} w_z \rho_{y,x|Z=z}}{\sum_{z=1}^{L} w_z} \tag{5.7}$$

where L is the number of groups and w_z is the weight given for group Z. For the purposes of this dissertation, it is assumed that $w_z = 1$. In applications different from the current one, these weights might be chosen to be unequal, to favor some groups over others. Let $\underline{X}^{(z)} = (x_i : i \in J_z)$ where J_z is the subset of values in group Z. Then ρ in this case is defined as:

$$\rho_{y,x|Z=z} = \frac{\sum_{i\in J_z} (R(X_i, \underline{X}^{(z)}) - \overline{R(X_i, \underline{X}^{(z)})})(R(Y_i, \underline{Y}^{(z)}) - \overline{R(Y_i, \underline{Y}^{(z)})})}{\sqrt{\sum_{i\in J_z} (R(X_i, \underline{X}^{(z)}) - \overline{R(X_i, \underline{X}^{(z)})})^2 \sum_{i\in J_z} (R(Y_i, \underline{Y}^{(z)}) - \overline{R(Y_i, \underline{Y}^{(z)})})^2}}. \tag{5.8}$$

The numerator of equation (5.8) can be written as

$$\sum_z \left(R(X_i, \underline{X}^{(z)}) - \overline{R(X_i, \underline{X}^{(z)})}\right)\left(R(Y_i, \underline{Y}^{(z)}) - \overline{R(Y_i, \underline{Y}^{(z)})}\right)$$
$$= \sum_z \left[\sum_{i=1}^{n_z} R(X_i, \underline{X}^{(z)}) R(Y_i, \underline{Y}^{(z)}) - \frac{n_z(n_z+1)^2}{4}\right].$$

Similarly the denominator can be simplified as follows:

$$\sum_z \left[R(X_i, \underline{X}^{(z)}) - \overline{(R(X_i, \underline{X}^{(z)}))^2}\right]$$
$$= \sum_z \left[\frac{n_z(n_z^2-1)}{12}\right].$$

Hence, the formula for S_2 can be rewritten as:

$$S_2 = \frac{1}{L}\sum_{z=1}^{L}\left(\frac{\sum_{i=1}^{n_z} R(X_i, \underline{X}^{(z)}) R(Y_i, \underline{Y}^{(z)}) - n_z(n_z+1)^2/4}{n_z(n_z^2-1)/12}\right) \tag{5.9}$$

To find the limiting distribution of S_2, it suffices to know the behavior of $\sum_{i=1}^{n_z} R(X_i, \underline{X}^{(z)}) R(Y_i, \underline{Y}^{(z)})$ for each Z.[1] Recall that from the Law of Large Numbers, it follows that

$$\frac{1}{n_z}\sum_{i=1}^{n_z} \frac{R(X_i, \underline{X}^{(z)})}{n_z}\frac{R(Y_i, \underline{Y}^{(z)})}{n_z} \approx E(F_{X_z}(X_z) F_{Y_z}(Y_z))$$

and

$$\sum_{i=1}^{n_z} R(X_i, \underline{X}^{(z)}) R(Y_i, \underline{Y}^{(z)}) \approx {n_z}^3 \int\int F_{X_z}(x_z) F_{Y_z}(y_z) f_{X_z,Y_z}(x_z, y_z) dx dy \tag{5.10}$$

Now,

$$F_{X_z}(x_z) \;=\; P(X_z \le x_z) = \Phi\left(\frac{x - \mu_{x_z}}{\sigma_{x_z}}\right).$$

[1]The random variables X and Y are superscripted with z to indicate that subsequent calculations are for one particular $Z - group$.

Similarly, $F_{Y_z}(y_z) = \Phi\left((y-\mu_{y_z})/\sigma_{y_z}\right)$. The joint distribution $f_{X_z,Y_z}(x_z, y_z)$ is bivariate normal given Z so

$$f_{X_z,Y_z}(x_z,y_z) = \frac{1}{2\pi\sqrt{1-\rho}} \exp\left\{ -\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_z-\mu_{x_z}}{\sigma_{x_z}}\right)^2 \right.\right.$$
$$\left.\left. -2\rho\left(\frac{x_z-\mu_{x_z}}{\sigma_{x_z}}\right)\left(\frac{y_z-\mu_{y_z}}{\sigma_{y_z}}\right) + \left(\frac{y_z-\mu_{y_z}}{\sigma_{y_z}}\right)^2 \right] \right\}.$$

This expression becomes

$$n_z^3 \int\int \Phi\left(\frac{x_z-\mu_{x_z}}{\sigma_{x_z}}\right) \cdot \Phi\left(\frac{y_z-\mu_{y_z}}{\sigma_{y_z}}\right) \cdot \frac{1}{2\pi\sqrt{1-\rho}}$$
$$\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_z-\mu_{x_z}}{\sigma_{x_z}}\right)^2 -2\rho\left(\frac{x_z-\mu_{x_z}}{\sigma_{x_z}}\right)\left(\frac{y_z-\mu_{y_z}}{\sigma_{y_z}}\right)+\left(\frac{y_z-\mu_{y_z}}{\sigma_{y_z}}\right)^2 \right]\right\} dx_z dy_z \tag{5.11}$$

This integral is given as in equation (5.5) where now $g(s,t^*) = \Phi(s)\Phi(s\rho_z + \sqrt{1-\rho_z^2}t^*)$ and the Gaussian quadrature formula in equation (5.6) again approximates it closely by choice of appropriate number of nodes. In this method, this formula is calculated for each Z and S_2 is the sum of the resulting values.

5.3 Simulation Study

Simulation results are presented in this section in order to study the behavior of the test statistics S_1 and S_2. Simulation provides the tools for examining the probabilistic behavior of these statistics for a wide variety of inputs and conditions. The simulation is done to compare the critical values and empirical power of the suggested rank statistics. The number of groups varied; so were the different magnitudes of within group variances and the patterns of within group correlations in the simulation. This analysis enables us to determine if one of the two statistics

systematically outperforms the other and to determine the better choice of rank correlation statistic, considering the selected factors.

5.3.1 Parameter Selection

The collection of parameters were designed for this study so that a wide range of possible conditions for the simulated data would be captured. The variances were chosen in three ways: variances in arithmetic progression scrambled in order (var1), variances somewhat asymmetrically placed (var2), and constant variances (var3) (see Table 5.1). All variances have been scaled so that the within group variances sum to one. For selection of ρ parameters, instances with varied directions and sign were included. Thus, a large variety of ρ vectors were tested in this research and I present results only for selected representative cases. For instance, (r1) is peaked in the middle and all the same sign, (r2) is monotone decreasing, and (r3) is a mix of negative and positive correlations (See Table 5.2). The magnitude of the within group correlations in each case are bounded by $2/\sqrt{n_z}$, $4/\sqrt{n_z}$, or $6/\sqrt{n_z}$, where n_z is the number of elements in each group and the number of groups, $L = 3$, 5, and 9. This sets the criteria for the hypothesis test as explained in the next section. The study test set was fully crossed with the 3×3 choices of variance and ρ vectors. However, the distribution choice (Normal), mean vector ($\mu_z = 0$), and proportion of samples in each group Z ($p_z = 1/N$) remained constant across cases.

Table 5.1: Group-wise Variance Parameters

$L = 3$	var1	var2	var3
	.333	.304	.333
	.444	.443	.333
	.222	.253	.333
$L = 5$	var1	var2	var3
	.176	.120	.20
	.059	.287	.20
	.117	.185	.20
	.294	.167	.20
	.353	.241	.20
$L = 9$	var1	var2	var3
	.067	.068	.11
	.156	.083	.11
	.044	.182	.11
	.178	.156	.11
	.022	.141	.11
	.011	.109	.11
	.200	.052	.11
	.133	.114	.11
	.089	.094	.11

5.3.2 Simulation and Two-sided Test Procedure

The null hypothesis of interest in this study can be expressed as a test for no correlational relationship written as H_0 : all correlations are 0. The alternative hypothesis is $H_1 : r_z = a_z/\sqrt{n_z}$, where a_z represents some constant value (2, 4, or 6 in this dissertation) denoting magnitude of nonzero correlation in units of $1/\sqrt{n_z}$. These contiguous alternatives (very close to the null value of zero) [40] were tested to achieve limiting large-sample powers strictly between the α levels (.05 and .10) and 1.

The Monte Carlo simulation to conduct this test involves several steps. First, $B = 500$ samples of size $N = 1000$ indexed by $b = 1, ldots, B$, were generated for

Table 5.2: Group-wise Correlation Parameters bounded by (i) $2/\sqrt{n_z}$, (ii) $4/\sqrt{n_z}$, and (iii) $6/\sqrt{n_z}$ for $L = 3$, 5, and 9, where $n_z = 1000/L$.

L=3	r1	r2	r3		r1	r2	r3		r1	r2	r3
(i)	.050	.100	.060	(ii)	.150	.200	.260	(iii)	.250	.290	.360
	.110	.040	-.070		.220	.140	-.170		.330	.160	-.230
	.090	.008	.120		.190	.018	.120		.090	.088	.120
L=5	r1	r2	r3		r1	r2	r3		r1	r2	r3
	.007	.160	.001		.017	.310	.001		.210	.460	.100
	.050	.070	-.050		.250	.270	-.200		340	.320	-.050
	.140	.020	-.019		.280	.120	-.110		.420	.160	-.419
	.090	.016	.090		.190	.060	.290		.190	.080	.270
	.009	.003	.120		.009	.003	-.220		.090	.015	-.320
L=9	r1	r2	r3		r1	r2	r3		r1	r2	r3
	.003	.180	.020		.030	.400	.220		.130	.580	.020
	.023	.130	-.080		.123	.330	-.180		.223	.430	-.380
	.090	.109	-.190		.290	.209	-.190		.390	.309	-.090
	.140	.095	0.130		.340	.165	.330		.440	.250	.530
	.190	.060	.010		.380	.120	.410		.570	.160	.410
	.100	.052	-.025		.200	.060	-.225		.400	.082	-.136
	.070	.043	0.200		.170	.043	.200		.270	.043	.320
	.052	.026	-.037		.050	.026	.137		.152	.016	-.370
	.006	.008	.004		.006	.008.	.240		.060	.008	.040

each variance and ρ parameter combination. Next, the test statistics S_1 and S_2 was calculated theoretically, by using the asymptotic formulas in sections 5.1 and 5.2, as well as empirically by computing the average statistic across Monte Carlo replications. This generated the alternative distribution. For every fifth value of b, $R = 750$ random (within group) permutations were performed, and the same statistics were recalculated for each null hypothesis permutation.[2] Lastly, empirical power was calculated as the proportion of the 750 permuted samples of the test statistic under the alternative hypothesis that exceeded the significance threshold or $P_1(|S| \geq c)$. This cut-off threshold (c) was calculated from the empirical distribution

[2]In earlier tests, computed values were shown to be pretty stable so I chose to select every fifth B for the final calculations to save on computation time.

under the null hypothesis at each α.

5.3.3 Theoretical Formulas Compared to Empirical Averages

First, the empirical distributions of the test statistics obtained from the simulation were explored. When comparing the computations for the two statistics, the results for nearby alternatives when $L = 3$, 5, and 9 are presented in Appendix A. In general, the agreement between simulated and theoretical expectations is very high for both statistics with a difference of no more than two empirical standard errors. The results confirm that the expected values for S_2 do not depend on the variance parameter combinations and hence remain constant throughout. Furthermore, in the case where the within group variances are constant, it is found that either statistic can be applied because the averages of S_1 and S_2 are approximately the same and the results are indistinguishable. In Tables A.1 through A.3 it is observed that for $L = 3$, S_1 and S_2 are comparable in magnitude for all rho parameterizations except in the mixed-sign case (r3). In fact, there is less difference in S_1 and S_2 throughout for smaller numbers of levels of Z. In the cases where $L = 5$ and $L = 9$, S_2 almost always yields a systematically larger rank correlation than those from S_1.

5.3.4 Empirical Power Results

The critical values simulated from the null distribution for the statistics S_1 and S_2 for typical levels of significance ($\alpha = .05$ and .10) are provided in Table 5.3 for each α and Z-level. The critical values for S_1 are fairly similar for each

value of Z and in all cases, the critical values of S_2 are slightly lower than those of S_1. Results, using these cut-offs, of the empirical statistical power of the test for each two-sided test procedure under consideration are shown in Appendix B for Z-group and significance level. The test with the highest power for a given level of significance is always preferred.

Table 5.3: Simulated Critical Values for S_1 and S_2 for $\alpha = .05$ and .10.

	S_1		S_2	
	α			
	.05	.10	.05	.10
L=3	.064	.054	.062	.052
L=5	.069	.058	.062	.052
L=9	.069	.058	.062	.052

A comparison of the power of the tests indicates that for both statistics the test attains highest power for the case where within group correlation is peaked in the middle and all the same sign (r1) and lowest power when the correlations are mixed in sign (r3). There is more of a distinction of this in the case where $L = 3$. In general, statistic S_2 tends to achieve higher power and this is more evident for cases where there are more groups (e.g., 5 or 9) and closer alternatives (i.e., correlations are bounded by $2/\sqrt{n_z}$).

5.4 Comparison of Empirical Power to Normal Distribution

Inspection of the distribution of statistics S_1 and S_2 in the simulation reveals that these statistics are approximately normally distributed under both the null and alternative hypotheses. This is shown via histograms for select parameter choices

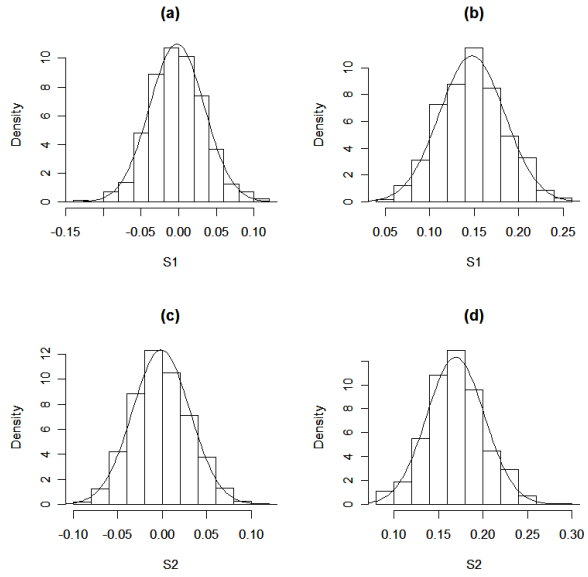


Figure 5.1: Histogram of S_1 and S_2 for $L = 3$ with parameter choices: variance = Var3, rho = r3, and contiguous alternative $a = 6$. Normal density overlaid for null ((a),(c)) and alternative ((b),(d)) hypotheses.

in Figures 5.1, 5.2, and 5.3. In each case, the normal $N(\mu_0, \sigma_0)$ and $N(\mu_1, \sigma_1)$ density functions for the null and alternative distributions, respectively are overlaid. Consequently, the empirical power results from section 5.3.4 can be compared with the normal distribution rejection probability.

Under the normal distribution, the power against the two-sided hypothesis for my test statistics becomes $P_1(|S| \geq \mu_0 + z_{\alpha/2}\sigma_0) = 1 - \Phi\left((\mu_0 - \mu_1 + z_{\alpha/2}\sigma_0)/\sigma_1\right) + \Phi\left((\mu_0 - \mu_1 - z_{\alpha/2}\sigma_0)/\sigma_1\right)$, where $\mu_0 = 0$. Tables in Appendix C show the power of statistics S_1 and S_2 re-calculated using the normal critical values $z_{\alpha/2} = 1.645$ and 1.96 for $\alpha = .10$ and .05, respectively. The power calculations of this form generally track very well with the previous calculations resulting in the Appendix B tables. This finding is significant in that permutational cut-offs are very costly in terms of computation time. In the future, the normal approximations for power for the test

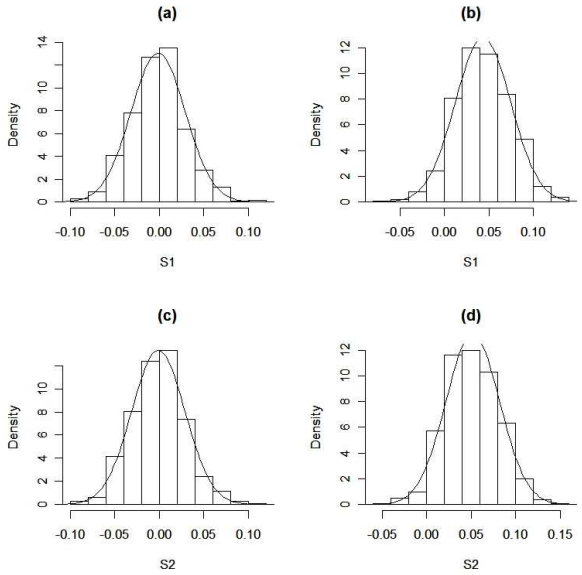


Figure 5.2: Histogram of S_1 and S_2 for $L = 5$ with parameter choices: variance = Var2, rho = r2, and contiguous alternative $a = 2$. Normal density overlaid for null ((a),(c)) and alternative ((b),(d)) hypotheses.

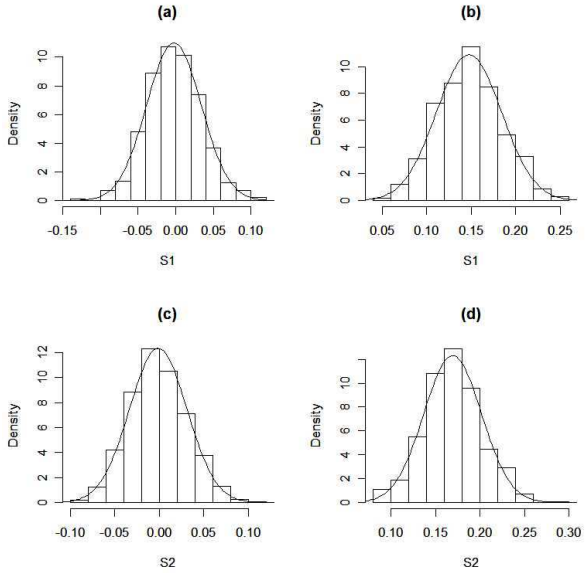


Figure 5.3: Histogram of S_1 and S_2 for $L = 9$ with parameter choices: variance = Var1, rho = r1, and contiguous alternative $a = 4$. Normal density overlaid for null ((a),(c)) and alternative ((b),(d)) hypotheses.

statistics can be used.

5.5 Summary

In Chapter 4, the partial rank correlation statistics S_1 and S_2 are used to summarize the relationship between automated Machine Translation metrics and task performance on the WH-extraction task. The permutational cut-offs in the final analyses suggested that the within group correlation structure did remain for the MT data for both statistics. This chapter explored these two statistics used under several distributional assumptions in a manner that enables the determination of the statistic that serves as a better choice for this comparison asymptotically. The results show that in general, statistic S_2 appears to yield a higher power for detecting differences in within group correlation. Chapter 6 uses the knowledge that an inherent relationship exists between the metrics and task performance. The magnitude of the partial rank correlation for each group (MT, WH, or MT X WH) and autometric (BLEU, METEOR, GTM, and oTER) serve as indicators of factors that make good predictors in modeling this relationship.

Chapter 6

Data Analysis Phase 2: Model Building, Evaluation, and Results

In Chapter 4, a correlation analysis was applied to establish that autometrics are indeed associated with subject task performance in the extraction task. This chapter extends the research beyond correlations and investigates the use of statistical modelling techniques to identify and characterize the *predictive* relationship between machine translated document quality, as judged by the four automated measures studied, and the outcome of the extraction task performed by subjects using the same collection of translated documents. Logistic regression, a convenient instance of a binary response Generalized Linear Model [47], was chosen to develop relationships between presence/absence of correct matches from subject responses in this experiment and autometric document-level scores.

This chapter presents general background and discusses the motivation for selecting this class of models. The specification, fitting, and interpretation of generalized linear models describing the dependence of subject performance on the extraction task of autometrics and other document features is summarized. Best

fitting fixed effect models show that autometrics are useful in distinguishing task-based performance of MT engines and under specific response criteria, and that certain MT engines do outperform others on subject responses for the extraction task. The consequences of such models, their effectiveness, statistical adequacy, and limitations as predictive tools are addressed. The modelling results are analyzed to determine which autometric or class of autometrics is more useful in predicting document utility and outline how interpretations of the models will enable us in future Machine Translation evaluations.

6.1 Statistical Modelling

Statistical modelling can be used to summarize experimental data. Model fitting involves the following steps [18]:

- Model specification – the mathematical formula chosen to relate the expected response (hit rate) to the explanatory variables, and the probability distribution of the response.

- Parameter estimation – the method of estimation (maximum likelihood in this dissertation), including the numerical algorithm and output summary statistics, which are all part of standard statistical software packages for widely used models like generalized linear models.

- Model validation – the statistics and diagnostic exhibits to be used in measuring the fit of each model to the data.

- Inference – confirmatory conclusions and interpretations that can be drawn about the model governing the data.

The sequences of these steps related to this work are discussed in the remainder of this section.

6.1.1 Logistic Regression

For the generalized linear model the expected response variable $\mu_i = E(Y_i)$ for the i'th observation or case is specified as a known function of a linear combination

$$\eta_i = \alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \tag{6.1}$$

of the explanatory variables $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ in the i'th case, where the coefficients α (called *intercept*) and β are unknown parameters, common to all observations and estimated from the data. The GLM requires further specification of a known *link function* g to capture the relationship between $E(Y_i)$ and $\mathbf{x}_i$, in the form

$$g(\mu_i) = \eta_i.$$

The modeler chooses the function g as well as the form of the probability distribution of the response Y_i (conditionally given η_i), from a so-called *exponential distributional family* [1]. One appropriate choice, associated with the particular GLM called the *Logistic Regression Model*, is called the *logit* link function where $\mu_i = E(Y_i)$ satisfies $\eta_i = g(\mu_i) = \log(\mu_i/(1-\mu_i))$ along with the *Binomial* distri-

butional family, and implies the relationship

$$E(Y_i) \;=\; \frac{\exp(\alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}{1 + \exp(\alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}. \tag{6.2}$$

The logistic regression model, like all GLM's, expresses the predicted or expected behavior of the response variable Y_i through an explicit parameterization (6.2) in terms of the explanatory variables. The number p of predictive terms entering into the model is often changed or updated after looking at the data during model fitting. The parameters are readily estimated from data using the method of maximum likelihood, a standard component of most statistical packages. There are several methods that can be used to assess model fit. A few techniques are discussed in Section 6.2. All statistical modelling analyses were carried out with **R**, a free variant of the S-plus statistical software package [53]. More statistical background for model-building is the focus of selected chapters in [1].

6.1.2 Logistic Regression for MT Evaluation

The data used for this research has several characteristics that makes logistic regression an appropriate choice of models; for example, the subjects' answer is either correct or incorrect and logistic regression specifies the probability of an event occurring or not occurring. The data also contains both categorical (qualitative or discrete) and continuous variables which can be easily handled simultaneously with logistic regression. Logistic regression then models the logarithm of the odds of a success in the extraction task.

By using statistical modelling, I am seeking to detect coefficients within logistic

regression models which are *statistically significantly* different from 0. This will help indicate whether specific predictor variables and their interactions (described in the next section) have an influence on the responses collected in the experiment. MT system, WH-type, and variables such as autometrics are related to a probability p through the function g in (6.1), where $g(p) = log(\frac{p}{1-p})$.

6.1.3 Model Variables

The goal of any analysis using regression is to find the best-fitting and most parsimonious, yet reasonable model to describe the relationship between the outcome (response) variable and a set of independent (predictor or explanatory) variables [30]. In the extraction experiment described in Chapter 3, several variables that could potentially summarize the data.

Response Variables

The design of the extraction task allowed subjects to identify specific WH-type items from each machine translated document. The response is defined as a success (that is, correct subject mark) or failure (that is, incorrect subject mark). This analysis focuses on the *hits* response rate (**hit rate**) as the response variable of interest. Recall from Chapter 3 that a 'hit' is the event that a subject selected a correct item (as validated by the reference truth). In this setting, $Y_i = \texttt{Hits}_i/\texttt{RTMtot}_i$ is modeled as though $\texttt{RTMtot}_i$, $\mathbf{x}_i$ were exogenous random inputs, conditioned on and thereafter fixed in modelling the conditional distribution of y given these variables; the number

$\texttt{Hits}_i$ of hits is treated as the number of successes in $\texttt{RTMtot}_i$ independent coin tosses[1], each with success probability given by $p_i = E(Y_i)$ satisfying $g(p_i) = \log(p_i\,/(1-p_i)\,)$. Other event rates collected in the study may also be modeled. If one were interested in assessing overall performance of subjects using the three-way possibility of response choices (correct response, incorrect response, and non-response), this could be accomplished using a regression setting that enables prediction of polytomous responses or by predicting each response separately and using a suitable function to combine the results of each individual prediction. Such models might also incorporate shared parameters across the probabilities for different responses.

Predictor Variables

Continuous Variables

Studying the relationship between automated metric scores and task performance is of most interest in this study. These metrics provide a sensible way to assess document quality and are readily accessible. Thus, if they can be shown to predict task ability, this will provide great insight for developers and users of MT systems. The scores with suffix 'avg', as introduced in Section 4.5, 'averaged' metric scores across documents and MT systems and attempted to get a document difficulty metric for each document regardless of the system that translated it. These scores were

[1]Using the RTMtot value strictly as a number of independent trials and conditioning on it was a modelling choice. The trials might not be independent, and the proper link might depend upon the RTMtot value. Although this possibility seems unlikely to affect the predictive properties of the models studied, it has not been investigated fully.

introduced to create a metric that could be more successful in drawing conclusions about documents for difficulty or *translatability*. Translatability refers to attributes of document quality. This concept of quality determines whether a document is easy or hard on a given scale of difficulty. Thus, a system could be evaluated based on how well it translated documents of a specific level and also based on how well certain tasks are performed in relation to the documents' degree of difficulty. This exploits the notion of quality vs. usefulness. The variables that serve as quantitative predictors in the models are:

- **BLEU and BLEUavg**
- **GTM and GTMavg**
- **METEOR and METEORavg**
- **oTER and oTERavg**

Categorical Variables

The distinction between documents of varying WH-type and MT systems was made with the intention of addressing the question of whether subjects perform better on extracting certain WH-types or output produced by certain systems. Although the autometric/task-performance relationship is what was ultimately sought, it is known from the work in Chapter 4 that document WH and MT classifications are important in summarizing this relationship. Thus, the categorical variables in the models are:

- **WH Type**– *Who, When, Where*
- **MT System**– *MT-1, MT-2, MT-3*

These categorical variables in the models are represented by what are called *dummy* variables for WH-categories and MT systems. If a specific variable has k levels, then the dummy or indicator for the j'th level (for j ranging from 1 to k) has value 1 if the categorical variable for a specific data-record has value j, and is 0 otherwise. For instance, the variable for MT-2 could be coded as $I_{[MT=2]}$ and would be valued at 1 when relating to a particular instance of an item from translation system 2. In general, if a nominal scaled variable has k possible values, then k-1 dummy variables will be needed [30] to describe the main effect of the categorical variable. In this case, an example model including MT variables would be

$$\log(\frac{p}{1-p}) = \alpha + \beta_1 I_{[MT=2]} + \beta_2 I_{[MT=3]}$$

and both dummy variables would be set to 0 in records corresponding to MT-1-translated documents.

Interaction Terms

The effect of a single factor in a model is called a *main effect* [48]. The variables described above can enter the models not only as main effects, but two or more factors can have a combined effect called an *interaction effect*. This happens when the effect of one of the variables is not constant over levels of the other. For example, an interaction between WH and BLEU implies that the coefficient for BLEU is different for WHEN,

WHERE, and WHO categories. Interaction terms are generally represented as predictor variables coded as products of quantitative predictors or of dummy variables for categorical predictors. In this work, an example model including interaction effects between WH variables and BLEU scores would be

$$\log(\frac{p}{1-p}) = \alpha + \beta_1 I_{[WH=WHERE]} + \beta_2 I_{[WH=WHO]} + \beta_3 BLEU$$
$$+\beta_4 I_{[WH=WHERE]} * BLEU + \beta_5 I_{[WH=WHO]} * BLEU.$$

According to [30], any interaction term included in a model should be based both on statistical considerations (i.e., if the added term significantly increases the model's fit) as well as practical considerations (i.e., the interaction must make sense from the study application).

6.2 Model Evaluation

The generalized linear modelling (GLM) framework helps to summarize the variation in the data and determine if response can be attributed to different characteristics of the data. Once the models have been fitted, it is necessary to determine how effective the models are at predicting the response variable. The validity or 'adequacy' of a statistical model, with respect to a dataset, is essentially the property that the observed data deviate from predicted or expected values by no more than would occur by chance or with large probability if the event-occurrence data were actually generated from the given set of predictor variables via the postulated model.

There exists an established body of statistical theory and "goodness of fit tests" [1] to assess the deviations between observations and predictions from fitted models, taking into account that the fitting of parameters was (or was not) done on the same dataset being used to test adequacy. Statistical models passing such tests of adequacy are the gold standard for applied statistical investigations, because most theoretically based statistical statements (for example, about the uncertainty in an estimated quantity) depend on using a 'correct' model in this sense. To evaluate the performance of models, several model-checking strategies are employed:

Chi-squared (χ^2) test [48] – measures goodness of fit of models to the data by measuring discrepancies between the observed values(o) of the data set and those predicted(e) by the model. This measure is relative to a specific categorization of the data into cells. In this chapter, the 9-cell MT × WH categorization as well as a 54-cell Document × MT categorization (53 after deletion of one document) are used. The formula for obtaining the (χ^2) statistic is

$$\sum \frac{(o-e)^2}{e}. \tag{6.3}$$

Likelihood Ratio Test (LRT) [1]– compares a constrained model 1 with predictors of interest fixed at nominal values to an unconstrained model 2 without those predictors in the form

$$\Lambda = -2\left[\frac{\text{max log likelihood of Model 2}}{\text{max log likelihood of Model 1}}\right] \tag{6.4}$$

where the 'likelihood' for discrete response data (as here) is the probability that the observed values of the response would occur with the observed values

of the explanatory variables. It is required that the models compared by the LRT approach be nested. The LRT statistic, under some regularity conditions which are satisfied here, approximately follows a chi-square distribution under the Model 1 probability law when data samples are large. Hence, to determine if the difference in likelihood between two models is statistically significant, the LRT is compared to a χ^2_p critical value from a standard statistical table.[2]

Akaike's Information Criterion (AIC) [1] –ranks a series of models by selecting a good model in terms of estimating quantities of interest rather than through significance tests. AIC penalizes a likelihood-based criterion of goodness of fit ("deviance") by the number of model parameters. Thus, the model that produces the smallest AIC is selected as the most likely representation of the given data. The AIC penalty helps offset the apparent increase in model performance that is attributed simply to a larger number of predictor variables. References [2] and [1] describe this criterion that selects a model as one that minimizes

$$AIC = -2[\text{maximized log likelihood} - \text{number of degrees of freedom}]. \quad (6.5)$$

6.3 Univariate Models

Since the ultimate idea is to describe the probability of "success" on the WH-extraction task as a function of a single autometric score, as an initial screen, a series of univariate models of the form (6.6) analyzing each predictor separately

[2] p is the degrees of freedom (difference in parameter dimension between the two models).

were tested.[3]

$$\textbf{Model 1} \qquad log(\frac{pHits}{1-pHits}) \quad = \quad \alpha \quad + \quad \beta_1 \quad * \quad MetricScore \quad (6.6)$$

The results for each of the univariate models for autometrics and their re-codes are displayed in Table 6.1. The p-value in the fifth column of the table is associated with the test applying the Wald Statistic, $(\hat{\beta}/StdErr(\beta))^2$. This comparison of the model parameter estimate to its standard error is a standard procedure for determining significance of coefficients of parameters in a model [30] and is compared to a chi-square distribution with 1 degree of freedom. The chi-square statistic in the final column refers to a different hypothesis and corresponds to the test of fit (observed vs. predicted) with respect to the 9 MT $\times$ WH cells.[4] The null hypothesis for the test on the chi-square statistic here is that the numbers of correct extractions (hits) follow the logistic regression model. The test statistic is compared with a χ^2 distribution with the degrees of freedom equal to the number of cells (9, in this work) minus the number of parameters in the logistic regression model. In these univariate models, the number of parameters is 1 for the intercept plus 1 for the metric so the degrees of freedom are 7.

The estimated coefficient in such a univariate logistic model is interpreted as the increment in the linear score due to a unit increase in the predictor variable. For example, in the METEOR only model, the term METEOR results in a change in

[3]The steps in this section are analogous to the single-variable correlations (between hit rates and metrics). The univariate models are presented here only to demonstrate in the logistic regression setting the ability or inability of individual autometrics to predict hit rates.

[4]As a disclaimer, these should be interpreted with caution because of the assumptions which would be needed to justify them, essentially that the translated- document-by-subject cases fall independently subject to a fixed array of probabilities within MT $\times$ WH cells.

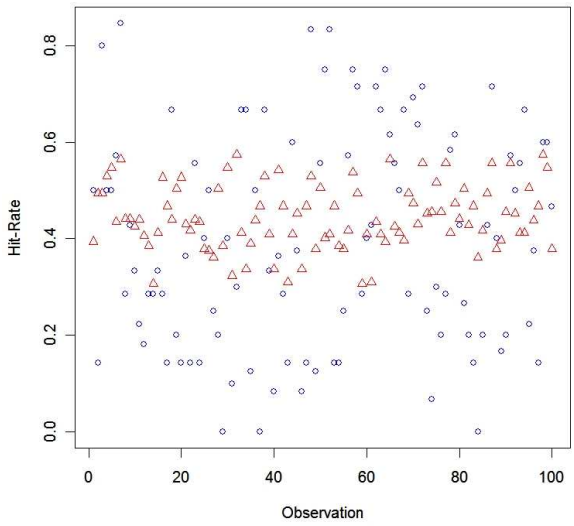
Table 6.1: Logistic Regression results for Univariate Models including estimated coefficients for each model along with their standard errors and significance as determined by the Wald Statistic p-value and chi-square statistic values for observed vs. predicted with respect to MT × WH.

Model	Variable	Est. Coeff.	Std. Error	p-value	χ^2_1
BLEU	Intercept	-.534	.04	<2e-16	449.61
	BLEU	2.77	.28	<2e-16	
GTM	Intercept	-1.52	.12	<2e-16	437.69
	GTM	2.44	.22	<2e-16	
Meteor	Intercept	-1.44	.10	<2e-16	410.85
	Meteor	2.79	.22	<2e-16	
oTER	Intercept	-.952	.07	<2e-16	432.13
	oTER	2.61	.24	<2e-16	
BLEUavg	Intercept	-.815	.07	<2e-16	469.32
	BLEUavg	.584	.07	<2e-16	
GTMavg	Intercept	-1.971	.22	<2e-16	481.27
	GTMavg	1.75	.22	2.00e-15	
METEORavg	Intercept	-.950	.19	3.56e-07	498.91
	METEORavg	.7119	.18	.0001	
oTERavg	Intercept	-.589	.13	9.5e-06	505.67
	oTERavg	.357	.13	.007	

$log(\frac{pHits}{1-pHits})$, that is, log-odds of correct task performance, of 2.79 over the intercept only model. All estimated coefficients in the displayed models are found to be extremely statistically significant, but none of the single-predictor models works well in predicting hit rates, according to the goodness of fit chi-square in the final column of Table 6.1. Although the univariate model with only autometric main effects would be better for future model interpretation by the MT community, Table 6.1 shows that even the 'best-fit' model at this stage, the model including METEOR only, has an extreme chi-square value of 27.94 on 7 degrees of freedom.[5]Figure 6.1 confirms the lack of fit of the autometric only models for 100 randomly selected individual

[5]The extreme p-values in Table 6.1 have no meaning other than as descriptive statistics since none of the models are statistically adequate.

Figure 6.1: Observed(circles) vs. METEOR-predicted(triangles) Hit rates for 100 random values



document cases of observed versus predicted values for METEOR. This indicates that more predictors are needed to account for task performance. Hence, one can not deduce from a document's metric score alone how well subjects will perform on the extraction task. The finding here confirms that not only does the strength of the relationships (discussed in Chapter 4) depend on grouping effects of the data, but so do any possible predictions.

6.3.1 Model Selection

The previous models, as well as the results in Chapter 4, lead us to believe that other variables in the data set must be useful in describing the hit rate. Several combinations of main effects variables and their interactions were tested in different models through a stepwise selection procedure. This approach to model building

is standard and can be implemented in most statistical packages ([1],[53]). Since the predictors described in Section 6.1.3 are all of interest in determining task performance, a model was tested with all of these variables as effects, as well as one with the interactions added. The most richly parameterized models yielded large standard errors for the estimated coefficients of interaction terms in these models. The interaction terms were highly correlated (multi-collinear) with the corresponding main-effect terms in the regression equation, making it impossible to disentangle the relative importance of main- and interaction- effects.

Keeping this in mind, models which support a significant reduction in deviance as judged by Likelihood Ratio Tests for models with and without parameters of interest, which have coefficients that are significant as tested by Wald statistics for significance of coefficients entering a model, and which did well in an external non-model based comparison (i.e. χ^2 of observed vs. expected value) were retained. Results of the best models fitting these criteria for each autometric are presented in the next section. The best and most parsimonious models include METEOR or oTER as predictors. However, if one includes more parameters and allows re-codes, then BLEUavg and GTMavg fare better. How I arrived at these conclusions is discussed in the next section.

6.4 Higher Dimensional Models for Each Autometric

After the series of models tested with single predictors (under-parameterized) and all possible predictors (over-parameterized), stepwise regression allowed us to

arrive at the best models associating respective autometrics with the probability of a person correctly marking a WH-item during the extraction task. The same selection procedure was followed separately for each autometric, allowing only that autometric and its 'avg' version to be considered for main effects and interactions with the categorical variables. Each model began with no predictors and a forward selection procedure was used to enter additional parameters, testing for significance of inclusion of each new variable at each stage. The chi-square for fit and the Wald chi-square for significance of the individual coefficients (in the presence of all the others) is displayed in Table 6.2.

In comparison to other models involving the respective autometrics, these four models (expressed in equation form in Models 2-5 below) proved to be the best models for each class of autometric based on AIC and the Wald statistic for individual term significance. Specific results regarding model fits are discussed in the next section.

Model 2

$$log(\frac{pHits}{1-pHits}) = -1.15 - .418 \times I_{[MT=2]} - .527 \times I_{[MT=3]} + 1.78 \times METEOR + 1.28 \times METEOR \times I_{[MT=2]} + 1.86 \times METEOR \times I_{[MT=3]} \quad (6.7)$$

Model 3

$$log(\frac{pHits}{1-pHits}) = -1.57 + .506 \times I_{[MT=2]} + .261 \times I_{[MT=3]} - .588 \times I_{[WH=WHERE]} - 3.60 \times I_{[WH=WHO]} + .960 \times GTMavg + .783 \times GTMavg \times I_{[WH=WHERE]} + 3.84 \times GTMavg \times I_{[WH=WHO]} \quad (6.8)$$

Table 6.2: Logistic Regression results for Best Higher Dimensional Models for each autometric including estimated coefficients for each model along with their Wald Statistic value and chi-square statistic values for observed vs. predicted with respect to the 9 MT × WH cells.

Model	Variable	Est. Coeff.	Wald χ^2	Num Param	$\chi^2_{9-numparam}$
Model 2	Intercept	-1.15	48.50	6	21.98
	MT2	-.418	1.89		
	MT3	-.527	3.53		
	METEOR	1.78	17.55		
	METEOR*MT2	1.28	3.93		
	METEOR*MT3	1.86	7.12		
Model 3	Intercept	-1.57	17.83	8	9.05
	MT2	.506	91.70		
	MT3	.261	24.98		
	WHERE	-.588	1.28		
	WHO	-3.60	34.71		
	GTMavg	.960	7.15		
	GTMavg*WHERE	.783	2.31		
	GTMavg*WHO	3.84	40.21		
Model 4	Intercept	-.771	45.04	8	8.31
	MT2	.508	91.89		
	MT3	.260	24.75		
	WHERE	-.064	.108		
	WHO	-1.20	49.72		
	BLEUavg	.177	3.10		
	BLEUavg*WHERE	.243	1.41		
	BLEUavg*WHO	1.23	65.24		
Model 5	Intercept	-1.30	70.16	6	14.76
	MT2	.658	5.09		
	MT3	.507	6.20		
	oTER	3.68	30.09		
	oTER*MT2	-1.89	4.10		
	oTER*MT3	-1.20	1.96		

Model 4

$$log(\frac{pHits}{1-pHits}) = -.771 + .508 \times I_{[MT=2]} + .260 \times I_{[MT=3]} - .064 \times I_{[WH=WHERE]} - 1.20 \times I_{[WH=WHO]} + .177 \times BLEUavg + .243 \times BLEUavg \times I_{[WH=WHERE]} + 1.23 \times BLEUavg \times I_{[WH=WHO]} \quad (6.9)$$

Model 5

$$log(\frac{pHits}{1-pHits}) = -1.30 + .658 \times I_{[MT=2]} + .507 \times I_{[MT=3]} + 3.68 \times oTER - 1.89 \times oTER \times I_{[MT=2]} - 1.20 \times oTER \times I_{[MT=3]} \quad (6.10)$$

A few points about these models must be clarified. In equations (6.8 and 6.9), two MT system and WH-type variables are represented. However, there are 3 of each under investigation in this study. The *R* software produces *n-1* linear predictors for each variable, where n is the number of factors of the variable. As mentioned in Section 6.1.3 describing model variables, the predictor in the model is interpreted as the contrast of the factor MT 2 against the omitted factor MT 1.

The Wald tests for the coefficients of the METEOR metric and the METEOR $\times$ MT-3 interaction indicate that they were the most significant variables in Model 2 as can be seen in Table 6.2. This means that documents with higher METEOR scores, and in particularly those from MT-3, have a significant effect on the change in hit rate. Similarly, in Models 3 and 4, with the exception of the WHERE variable and its interaction, all of the included variables and their interactions are highly significant. Lastly, in Model 5, higher oTER scores were most likely to result in correct matches because the oTER metric was the most significant predictor in the model.

Table 6.3: Observed vs Predicted Hits totaled over WH × MT

					Predicted Hits			
MT	WH	ObsHits	RTMTot	NumObs	Mod2	Mod3	Mod4	Mod5
1	WHEN	293	881	118	327.52	315.72	315.18	326.24
1	WHERE	428	1107	118	414.87	414.87	427.01	404.16
1	WHO	460	1103	118	438.61	437.85	438.80	450.60
2	WHEN	380	735	97	322.47	347.86	349.00	350.82
2	WHERE	563	1094	118	569.64	557.62	558.39	552.99
2	WHO	528	1097	118	578.89	565.53	563.60	567.19
3	WHEN	360	879	117	375.82	369.42	368.81	393.14
3	WHERE	489	1103	118	497.98	494.95	494.59	483.99
3	WHO	521	1104	118	496.19	505.63	506.59	492.87

6.4.1 Goodness of Fit for Higher Dimensional Models

Thus far, the successive stages of model fitting have led to the following parameters as best predictors with respect to each autometric:

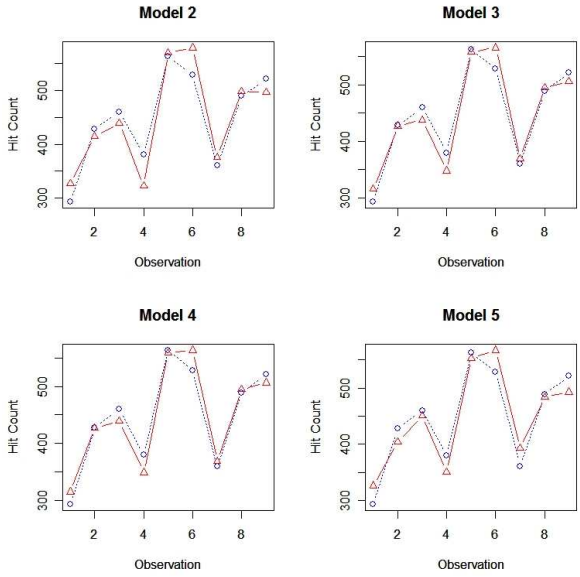
- Model 2: MT, METEOR, METEOR*MT
- Model 3: MT, WH, GTMavg, GTMavg*WH
- Model 4: MT, WH, BLEUavg, BLEUavg*WH
- Model 5: MT, oTER, oTER*MT

Models 2 and 5 have 6 parameters (intercept, 2 binary indicators for MT, 1 for metric, plus 2 for the metric*MT interaction) and Models 3 and 4 have 8 parameters (intercept, 4 binary indicators for MT and WH, 1 for metric, plus 2 for the metric*WH interaction). Table 6.3 displays the observed counts (found in the *Hits* column) vs. predicted or expected counts (found in columns *Mod2Hits-Mod5Hits*) tabulated across WH and MT. To check whether these discrepancies

are quantitatively greater than what might occur by chance, a chi-square test for the Hit rate was performed. The statistic was obtained as described in Section 6.2 over all cells of the table consisting of the 9 WH $\times$ MT cells. The chi-square constructed in this way for Models 3 and 4 is compared to the chi-square percentage point with 1 degree of freedom, calculated by the formula *number of cells - number of model parameters* (in this case, 9-8). These goodness-of-fit statistics show that Model 4 comes closest to adequately representing the proportions of hits in the data, with a chi-square value of approximately 8.31. Even though this value is extreme for a chi-square with 1 degree of freedom and shows Model 4 would not be a final model, it is still favorable considering the other model fitting stages. The chi-square values for models 2, 3, and 5 are 21.98, 9.05 ,and 14.76 on 3, 1, and 3 degrees of freedom respectively. Chi-square would have to be much smaller for the models to be judged 'statistically adequate', which may not be attainable without introducing other structure such as the between-subject variability into the model through random effects.

Figure 6.2 displays graphically the fit between the predicted values of the models in Table 6.3 and the observed data. Each model is very close to capturing the hit rate event counts, as seen by comparing its predicted counts to the observed counts by (WH and MT) category. It is evident here that even though the chi-square values for Models 2 and 5 are more extreme than for Models 3 and 4, the distance between observed versus predicted is fairly similar across models. Thus, with fewer parameters and without having to re-code the metric, METEOR and oTER are comparable in terms of predicting performance.

Figure 6.2: Observed(circles) vs. Predicted(triangles) Hit Counts–WH by MT for Models 2-5. The x-axis represents the 9 MT × WH cells and the y-axis represents the hit count for each cell.



6.5 Combined Autometric Models

Several authors have suggested using combinations of autometrics to determine translation quality and have reported comparisons with human translations for these combinations [3, 25, 37, 45]. Table 6.4 shows the results for the univariate models of Section 6.3, as well as the 3 possible combinations of metrics (pairwise, three a time, and all metrics). The only variables in these models are the metric (B = BLEU, G = GTM, M = METEOR, T = oTER) combinations. Combining multiple metrics in this setting boosted the predictive power for all metrics except METEOR. For instance, the χ^2 over cells from the 9 MT × WH cells demonstrate that METEOR is as good a predictor by itself as it is when combined with and one of the other three metrics. In fact it is found that both of the poorer performing individual metrics

Table 6.4: Logistic Regression Results for Combinations of Autometrics Fitted to the Data; B = BLEU, G = GTM, M = METEOR, T = oTER

Model	Deviance	9 cell χ^2	df	53 cell χ^2	df
B	1982.89	47.01	7	449.61	51
G	1961.78	70.45	7	410.85	51
M	1915.08	27.94	7	437.69	51
T	1959.21	30.01	7	432.13	51
B, G	1958.59	65.66	6	435.87	50
B, M	1914.81	26.81	6	410.75	50
B, T	1956.45	34.84	6	431.10	50
G, T	1942.75	48.97	6	425.27	50
G, M	1906.95	37.51	6	405.42	50
M, T	1911.78	28.74	6	408.30	50
B, G, M	1898.91	34.53	5	401.56	49
B, G, T	1942.37	48.46	5	425.09	49
B, M, T	1908.88	24.38	5	406.77	49
G, M, T	1906.67	36.83	5	405.59	49
B, G, M, T	1896.53	31.46	4	400.04	48

BLEU and GTM perform considerably better when paired with better performing metrics.[6]

Logistic regression models are fit with various combinations of only autometrics as predictor variables. Stepwise regression is again used to add in MT and WH effects since from previous results in this dissertation these quantities are important. I started from the best combination models in Table 6.4, adding or subtracting MT and WH effects based on their individual term significance as defined by the Wald test and their contribution to reducing the deviance in the model. Deviance is defined as the difference in $-2 * \log(Likelihood)$ of the current model and $-2 * \log(Likelihood)$ of the saturated model where the saturated model is the model with the number of parameters equal to the sample size [1]. Models that included

[6]Degrees of freedom in the table are derived as number of cells (9 or 53) minus number of parameters fitted to each respective model.

Table 6.5: Deviance results for Best Combined Metric Logistic Regression Models accounting for MT and WH effects.

Model	Num Parameters	Deviance
MT + Meteor + GTM	5	1861.85
MT + WH + BLEUavg + Meteor	7	1861.75
MT + WH + METEORavg+BLEUavg+GTMavg	8	1761.15
WH + METEORavg + BLEUavg + GTMavg	6	1849.60
MT + WH + BLEU + oTER + Meteor	8	1851.64

Table 6.6: MT × WH and MT × WH × Rep χ^2 Goodness-of-Fit results for Best Combined Metric Logistic Regression Models accounting for MT and WH effects.

Model	9 cell χ^2	df	53 cell χ^2	df
MT + Meteor + GTM	25.16	4	388.64	4
MT + WH + BLEUavg + Meteor	17.52	2	400.77	2
MT + WH + METEORavg+BLEUavg+GTMavg	7.83	1	345.82	1
WH + METEORavg + BLEUavg + GTMavg	57.05	3	386.82	3
MT + WH + BLEU + oTER + Meteor	20.69	1	385.14	1

variables that decreased the deviance a significant amount in relation to the number of additional parameters were favored over those that did not.

The next step of model building involved checking if the re-coded *avg* metrics performed better than original metrics in combinations as in cases in earlier trials. The diagnostic results for the best competing models combining metrics including MT and WH effects are in Tables 6.5 and 6.6. The latter table displays goodness of fit results for both observed versus predicted in the 9 MT × WH cells and the 53 MT × WH × Rep cells.[7] The 'best' model found at this stage for the occurrence of Hits was:

[7]Recall that one outlier document was removed so there are now 53 versus 54 total machine translated documents.

Model 6

$$log(\frac{pHits}{1-pHits}) = .570 + .498 \times I_{[MT=2]} + .262 \times I_{[MT=3]} + .848 \times I_{[WH=WHERE]} + .741 \times I_{[WH=WHO]} + 1.67 \times BLEUavg - 4.78 \times METEORavg + 1.51 \times GTMavg \quad (6.11)$$

Recall that *MT = 1* and *WH = WHEN* respectively served as baseline categories for the MT and WH variables. Because there were significant contrasts between both MT systems MT-2 and MT-3 with MT-1 and both Wh-types Where and Who with When, these MT and WH predictors were included in the final model. In addition, the re-coded metrics BLEUavg, METEORavg, and GTMavg proved useful as predictors for hit rate and can be interpreted as a measure of the degree of difficulty of a translated document. Additionally, I found that the METEOR metric in its original form is a highly significant predictor of task performance at different stages of modelling.

A logical conclusion of the final Model 6 is that the contrast between MT-2 and MT-3 should be distinguished with respect to success on the information extraction task, as well as the WH-types Where and Who. Another interesting finding is that in this model, METEORavg has a negative coefficient. This is puzzling because all metrics have been coded individually so that at document level a higher metric score indicates a more linguistically successful translation. It is possible that the negative METEORavg coefficient is due to the overlapping components of these metrics, for example, they all include some form of precision and/or recall; and what Model 6 actually represents is the contrasts among the averaged metrics as

being of primary predictive value. I stop at this model because through all the stepwise regression procedures and statistical tests, this model show superior results in terms of parsimony in comparison to others found. χ^2 results presented in this section need to be smaller in order for these fixed effect models to be adequate. Later findings show random effects should be included to obtain statistical adequacy.

6.5.1 Modelling Summary

An effective model for predicting aggregated hit rates has been developed, even as the individual subject-marked translated document hit rates are intrinsically *not* predictable from these models. The comparisons between models show that models that include main effects for MT, WH, BLEUavg, METEOERavg, and GTMavg show success in reproducing overall hit rates within the 9 categories defined in Table 6.3. The most successful model found does not yet achieve full statistical adequacy but still promises to be very useful. In the next section, permutational methods similar to those in previous chapters are utilized to show how the predictors can be used in the model as true indicators of task performance based on the permutational significance of their coefficients.

It should be noted that a valid model might not be fully predictive in the sense that there might be sources of random variation from human subject-to-subject variability, which could not be modeled specifically. I tried treating each subject as its own individual variable in a model but this model is highly over-parameterized because there are 59 subjects. Several possible groupings were also examined based

on demographic information obtained from the subjects during the study, like job title, years of experience, or experience with MT systems, but none of these variables were closely correlated with subject-specific hit rates. Further study could consider a predictor for subjects in the final model. In this case, a *random- or mixed-effect* model might describe the resulting subject-specific increments to event-rate in terms of a random variable with a specified distribution for subject variations, such as a normally distributed variate with mean 0 and variance fitted as a model parameter.

6.6 Permutational Significance of Autometric Coefficients

When traditional approaches to determining relevance of models and individual coefficients are not adequate, it is unclear in the literature what strategies are best in providing some interpretation for the quality of the model in the absence of adequacy. Some empirical studies suggest creating a permutation distribution for tests of individual terms [5], although the exact methods are not settled upon. Permutation procedures similar to those described in Chapter 5 were used to test the significance of the β coefficients for individual autometrics in models after accounting for MT and WH groups. A permutation test in this manner calculates the probability of getting a Wald statistic value ($\hat{\beta}/StdErr(\beta)$) equal to or more extreme than the observed value of the test statistic in the original model. The autometric scores were randomly re-shuffled and re-fit the models with the random re-orderings. In all cases, the coefficients for autometrics in each model were found to be highly significant with a permutational p-value of less than .001. This result

Table 6.7: Number of Observed vs. Predicted Hits for Model 6 with respect to the 9 MT × WH cells

MT	WH	RTMTot	NumObs	ObsHits	Mod6Hits
1	WHEN	881	118	293	314.48
1	WHERE	1107	118	428	429.38
1	WHO	1103	118	460	437.14
2	WHEN	735	97	380	351.75
2	WHERE	1094	118	563	555.50
2	WHO	1097	118	528	563.75
3	WHEN	879	117	360	366.76
3	WHERE	1103	118	489	495.12
3	WHO	1104	118	521	508.11

establishes the significance of the autometric predictors in their respective models.

6.7 Cross-Validation Results

Now that a wide set of logistic regression models have been analyzed, it would be beneficial to know how well the best fitted model predicts unseen data, for example, what the model predicts about the performance of MT systems on an Information Extraction Task under different WH categories and document-difficulty scenarios. The ability of any of the models presented in this chapter to predict task performance would be the prediction accuracy. For example, Table 6.7 has the estimated number of correct marks in the task from Model 6 which yielded the smallest difference between the known or observed values and the estimated number of correct marks.

The different classes of models and different choices of predictors were compared within a cross-validation study in which approximately 80% of the experiment data (833 cases) was used as the *Training set* and the remaining 20% of the data

(207 cases) was withheld as the *Test set* to test the adequacy of the model produced from the training set. 70 cases were selected across each WH-type to include a balanced sample across who, where, when categories. This cross-validation was repeated by randomly sampling the test and training sets 1000 times from the data and computed the sum of squares error for each run.

The prediction accuracy for these models is estimated by the square root of the mean of the squared difference between the observed and predicted number of hits. The process used for calculating sum of squares error (SSE), mean square error (MSE), and root mean square error (RMSE) is as follows:

1. Compute sum of squares error on each run by using (test set) observed minus predicted (gotten from training) and place values into an SSE array.

2. Compute MSE at each run as the SSE/ 207 (207 is the total number of cases in the test set) and place values into an MSE array to produce 1000 MSEs calculated over all cases for each run.

3. Compute the average of the MSEs for each model and take the square root to get the RMSE.

Cross validation results for testing the overall error rate of models are highlighted in Table 6.8. The best prediction accuracy is given by the final model, Model 6 (in bold).

Table 6.8: MSE and RMSE values obtained from Cross-validation for Models 1-6

Model	MSE	RMSE
Mod1METEOR	3.70	1.92
Mod1BLEUavg	3.86	1.96
Mod1GTMavg	3.79	1.95
Mod1oTER	3.82	1.95
Mod2	3.71	1.93
Mod3	3.62	1.90
Mod4	3.54	1.88
Mod5	3.79	1.95
Mod6	**3.39**	**1.84**

6.8 Further Model Building with Random Effects

The best model until this point has been obtained by only considering parameters entering the model as fixed effects. Subsequent models fit with the inclusion of a variable for subject (which has 58 parameters) appear to produce a good fit to the response data and come closer to the observed hit rate.[8] However, treating each subject as its own individual effect is definitely an instance of overfitting because these models so are highly parameterized. Additionally, there are large standard errors amongst interaction terms making it difficult, if not impossible, to interpret what each effect means. Nevertheless, it would be worthwhile to consider a predictor for subjects in the final model. To address this issue, an approach using the subjects as random effects in a generalized linear mixed model is investigated. Although, adding in random effects do not help with statistical adequacy, the models do account for a considerable large amount of the variability that contributed to the lack of fit in fixed effect models.

[8]The number of parameters is obtained by the formula, Number of Subjects - 1

Table 6.9: Deviance Comparison between Fixed Effect Model 6 and Mixed Effects Models with Various Document Level and Subject Random Effects. Variance for each mixed effect is also given. MixMod3 has two random components, MT × Document and Subject, which are given respectively.

Models	Mod6	MixMod1	MixMod2	MixMod3
Deviance	1761.1	1493	1306	1176
Deviance Change	—	268.15	455.15	585.15
Random Effect Variance		.157	.302	.321/.121

6.8.1 Mixed Effect Models

Generalized linear mixed effect models (GLMMs) extend GLMS by adding random terms in the linear predictor to account for overdispersion [48]. The regression formula in Section 6.1.1 becomes

$$\log(\frac{p}{1-p}) = \alpha + \beta X + a + e \tag{6.12}$$

where a in this work represents one or more random error terms.

Several mixed effect models are tested with Model 6 as the base model. These models include (i) document cluster random effects (MixedMod1), (ii) MT × document cluster random effects (MixedMod2), and (iii) both MT cluster and Subject random effects (MixedMod3). Table 6.9 displays the results of model fitting through comparison between deviance of the fixed effect model (Model 6) to that of the model with random effects added at each level.

6.8.2 Random Effect Model Results

The fitted random effects models show that there are strong and significant document fixed-effects, beyond the "best model" (Model 6) described in Section

6.5, if the model keeps the fixed-effect terms in Model 6 and adds a document-level random intercept effect. Specifically, Model 6 already has 5 document-level fixed effects and deviance 1761.1 (as shown in Table 6.9) and MixedMod1 has deviance 1493. Results show that with the latter model, METEORavg and BLEUavg variables are the strongest individual metric related contributions to WH among the document-level fixed effects.

It is shown in Table 6.9 that the variance is progressively larger in each of the mixed effect models presented. The best model found with random effects is the MixedMod3 (deviance 1176 with 4 fixed effects and 2 random intercepts). This model starts from Model 6, deletes the GTMavg term, and incorporates a Doc-by-MT level normal random intercept (variance .321), as well as a very significant subject random effect (variance .121). In this final model, all of the fixed-effect coefficients other than MT-2 (which could also be deleted, indicating that the model finds no significant difference between MT systems 1 and 2) remain significant, but much less than either Model 6 or MixMod1.

6.9 Phase 2 Summary

I have examined in this chapter the predictive ability of autometrics in signifying WH-extraction task response rates. Best fixed-effect models found include recoded average autometrics for METEOR, BLEU, and GTM in the presence of MT and WH effects. These metrics demonstrated the most adequate goodness of fit statistics in comparison to others. I also show that although completely adequate

logistic regression models of this relationship were not found, these autometrics do give insight to correct task response as noted in the high significance of coefficients found permutationally in the models. This process gave us a basis for determining if one metric is permutationally more significant than others in predictions. Cross-validations on the performance of best models summarized the effectiveness [and limitations] of the models as a predictive tools. Lastly, I discussed the improvement in model results by introducing document by MT system and subject random effects. Under this scenario, the BLEUavg and METEORavg variables together are still the most important document level difficulty variables that describe task performance.

Chapter 7

Conclusion and Future Work

This dissertation has provided the first in-depth look at the connection between standard *measures of performance* of machine translation systems (assessed by automated machine translation evaluation metrics of quality) and task-based *measures of effectiveness* (assessed by task responses in a utility experiment). One goal of this study was to determine whether a relationship between automated machine translation evaluation metrics and task-based evaluation metrics exists. Another goal was to develop a predictive regression model of task performance to assess the effects of certain categorical (such as, MT system or WH type) or continuous (such as, BLEU score) characteristics of a translated document on subject performance on a well-defined task using those documents.

First, findings from an initial correlation analysis of the connection between these two MT evaluation paradigms were presented and contrasted with current strategies for evaluating translations. Next, a novel idea for assessing partial rank correlation within the presence of grouping factors was introduced. Lastly, a framework for task based machine translation (MT) evaluation and predictive modelling

of task responses was demonstrated. This was accomplished through an iterative approach to model building, testing, and fitting logistic regression models. The model building strategy gave new information about the relative predictive strengths of the different autometrics (and re-coded variants of them) within the statistical GLMs developed in analyses of the Information Extraction Task data. This work showed the lack of predictive ability of most current autometrics, as is, to predict task performance but showed that through the use of re-codes, near adequacy can be accomplished in a logistic regression setting. The rest of this chapter describes the contributions of this dissertation, limitations of the work presented here, and possible future research directions.

7.1 Contributions

The following contributions have been made by this dissertation:

- The investigation of the relationship between document quality and usefulness– Through a user-centered focus on translation evaluation using autometrics and an innovative analysis of data through applied statistical techniques, autometrics were used as tools for assessing translated documents.

- Correlation analysis on Automated Machine Translation metrics and Task-Based Responses from an Information Extraction Task—This work identified that although no one metric stood out in initial comparisons, there is a significant relationship between autometrics of translation quality as a whole and document utility in the extraction task.

- Validation of need to have more granularity with evaluation metrics—This dissertation demonstrated the need to utilize *document level* metrics for task handling purposes rather than *system level* approaches generally performed in the MT community. It also exposed the sensitivity to granularity when trying to assess task performance through correlations which are quite low. Permutational testing, however, showed that the relationships found were not ascribed to chance, thereby giving us motivation for going beyond the standard correlation analysis.

- Introduction and implementation of partial rank correlation statistics for assessing rank correlation in the presence of grouping factors—Methodolgical results are provided on characteristics of bivariate and partial rank correlations along with permutation tests of significance. Expressions were derived for partial rank correlation through two statistics and show that one of the statistics, a weighted combination of within group rank correlations, generally yielded a higher power for detecting within group correlations.

- Implementation of novel metric from re-coding autometrics—It was shown that a scheme for averaging autometrics across MT systems produced an overall *document difficulty* metric that performs well in indicating the probability of correct response on an extraction task.

- The first use of logistic regression to predict task performance using automated metrics as model predictors—Although METEOR appeared to be the best original metric to have the most interesting relationship with the task re-

sponse; the best predictor of task performance is found when recoded versions of METEOR, BLEU, and GTM are combined and coupled with MT and WH effects.

7.2 Limitations of the Study

The methods and techniques presented in this dissertation have the following limitations:

- Experiment Design—This study was the first of its kind so there are several lessons learned stemming from the design of the experiment. Future studies should address systems of more variable performance and documents yielding a larger range of metric scores. The systems chosen for this study, although different in structure, turned out to be very similar in terms of the task performance results and range of automated metrics on documents produced by each system. It is my belief that if there is a more defined scale of good (high) versus bad (low) translation quality scores and evident low, medium, and high performing MT systems (gained from prior knowledge of system performance), then relationships sought may be easier to tease apart and more applicable on a wider scale.

- Results for One Specific Task—Given that there are a myriad of uses for machine translated documents, there are several possible tasks that could have been chosen for this study. From prior pilot studies, I found that the information extraction task seemed to fall in the middle on the hierarchy of text

handling tasks as developed by [61]. Because one specific task was used in this study, the findings can not readily be carried over to other tasks of interest.

- Distribution of Partial Rank Statistics—In the methodological look at the behavior of the two partial rank statistics introduced, bivariate normality was used to reach limited conclusions via simulation about when one of the statistics S_1 versus S_2 was better than the other. For a more robust look at the asymptotic behavior of these statistics, it would be beneficial to consider other distributions and re-assess how the statistics behave.

7.3 Future Work

This section describes possible directions for future work arising from the research conducted in this dissertation.

Partial Rank Correlation Statistic: As discussed in the limitations section, in Chapter 5 the asymptotic behavior of partial rank correlation statistics was studied for a specific instance of bivariate normality. Further correction of statistics S_1 and S_2 by group-Z scale-factors as well as locations may make them generally more applicable in other settings. Further investigations by simulation about (relative) power will certainly be worthwhile future work, as will a search for other domains of applicability. Datasets which are highly cross-classified are often available in several applications such as ecology, biology, and the social sciences.

Assessment of Random Effect Models: Chapter 6 ends with the finding that Document by MT system random effects appear unavoidable in producing statistically adequate model descriptions. Assessment of statistical adequacy for GLMM extensions of the best models of this dissertation remains a topic for further research.

Leverage Autometric Features: This dissertation has described approaches that incorporate automated metrics as 'black boxes' in task-based evaluation. Each of these metrics have been built with specific features often trying to capture some aspect of translation quality not captured by its predecessor. For instance in Chapter 2, the BLEU metric is described as using a precision only approach to translation evaluation. Both GTM and METEOR add in recall as a factor with equal emphasis on precision and recall and more emphasis on recall, respectively. oTER explores an entirely different approach through the use of string edit distance. Given that there are obviously certain aspects of each metric that may or may not be beneficial in obtaining translation quality, it may prove valuable to study whether certain features or combination of features from autometrics offer better insight into task performance results.

Replication for Additional Tasks: Chapter 3 describes the full study in which I collected task performance results from subjects on three different text handling tasks. As noted in the limitations section, this dissertation only used response results from one of the tasks, information extraction. Although full-scale task-based experiments such as the one in this dissertation are expensive, more about

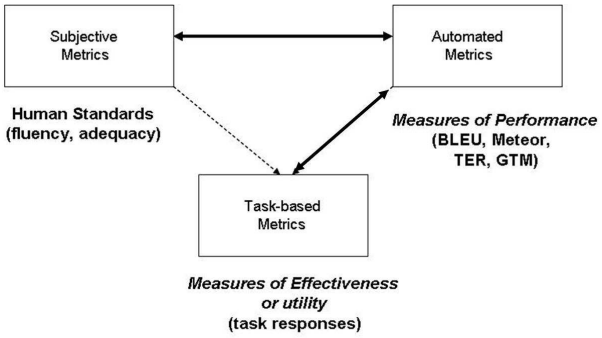
the relationship between other evaluation metrics and task performance would be found if the nature of this connection as it relates to other tasks was studied. I could either try to utilize the responses from the other two tasks (Categorization and Event Template-Filling) in the current experiment or replicate the study for additional tasks of interest. I would use my current results to leverage predictions about the other tasks and test the methods developed for more consensus.

Subjective Metrics: In Chapter 1, Figure 7.1 was used to introduce the progression of MT Evaluation paradigms. The diagram now has only one dashed line because the results in this dissertation offer insight to the previously un-studied autometric vs. task-based relationship. Because autometrics are the quickest and most cost-effective method for obtain translation quality results, it is apparent that the MT community will continue to use them in MT Evaluations. Furthermore, since comparison with human subjective judgments is the accepted way of validating autometrics, it is worth the user community investigating whether the results and methods presented in this dissertation can be used together with subjective judgments to obtain further improvements in MT Evaluation. This will provide a feedback loop between the three paradigms and determine the extent to which they complement each other.

7.4 Summary

In summary, this dissertation took a unique interdisciplinary look at a portion of the MT Evaluation problem for which the community currently has no consen-

Figure 7.1: Triangle of main MT evaluation paradigms. The bold line represents the work of past efforts, including the newly formed connection between Automated Metrics and Task-based Metrics as found in this dissertation. The dashed line represents a possibility for future work.



sus. This research accomplished this through the use of several applied statistical techniques and showed that model-building strategies in the context of GLM's are quite useful. Ultimately, MT evaluation methodology was extended to create new metrics specially relevant to task-based comparisons. Now users can begin to tie the intrinsic automated metrics to the extrinsic metrics for task they perform. The bottom-line was that there was need to average away MT dependence (averaged metrics performed better in overall predictions than original autometrics). Moreover, combinations of recoded metrics performed better than any individual metric.

Appendix A

Results Tables for Theoretical versus Empirical Values of S_1 and S_2

This section contains the tables described in Section 5.3.3 detailing the results for nearby alternatives when $L = 3$, 5, and 9.[1]

Table A.1: Theoretical and Empirical Values of S_1 and S_2 for each variance vector: (a) Var1, (b) Var2, (c) Var3 when $L = 3 and$a= 2. Standard error values are shown in parentheses.

	Rho	Theoretical		Average	
		S_1	S_2	S_1	S_2
(a)	r1	.080	.080	.080 (.001)	.079 (.001)
	r2	.050	.047	.050 (.001)	.047 (.001)
	r3	.016	.035	.016 (.001)	.035 (.001)
(b)	r1	.081	.080	.080 (.002)	.079 (.001)
	r2	.048	.047	.047 (.001)	.047 (.001)
	r3	.018	.035	.020 (.001)	.036 (.001)
(c)	r1	.080	.080	.079 (.001)	.079 (.001)
	r2	.047	.047	.045 (.001)	.045 (.001)
	r3	.035	.035	.033 (.001)	.033 (.001)

[1]All theoretical values are calculated using formulas (5.2) and (5.7) for S_1 and S_2 respectively.

Table A.2: Theoretical and Empirical Values of S_1 and S_2 for each variance vector: (a) Var1, (b) Var2, (c) Var3 when $L = 3 and a = 4$. Standard error values are shown in parentheses.

	Rho	Theoretical		Average	
		S_1	S_2	S_1	S_2
(a)	r1	.178	.179	.179 (.002)	.178 (.001)
	r2	.126	.114	.127 (.002)	.115 (.001)
	r3	.040	.067	.041 (.001)	.067 (.001)
(b)	r1	.180	.179	.181 (.001)	.180 (.001)
	r2	.120	.114	.120 (.001)	.113 (.001)
	r3	.036	.067	.038 (.001)	.069 (.001)
(c)	r1	.179	.179	.178 (.001)	.178 (.001)
	r2	.114	.114	.114 (.001)	.114 (.001)
	r3	.067	.067	.069 (.001)	.069 (.001)

Table A.3: Theoretical and Empirical Values of S_1 and S_2 for each variance vector: (a) Var1, (b) Var2, (c) Var3 when $L = 3 and a = 6$. Standard error values are shown in parentheses.

	Rho	Theoretical		Average	
		S_1	S_2	S_1	S_2
(a)	r1	.235	.214	.234 (.001)	.212 (.001)
	r2	.178	.172	.180 (.001)	.173 (.001)
	r3	.048	.080	.048 (.001)	.080 (.001)
(b)	r1	.231	.214	.228 (.001)	.211 (.001)
	r2	.172	.172	.172 (.001)	.171 (.001)
	r3	.041	.080	.044 (.001)	.083 (.002)
(c)	r1	.214	.214	.213 (.001)	.213 (.001)
	r2	.172	.172	.171 (.001)	.171 (.001)
	r3	.080	.080	.081 (.001)	.081 (.001)

Table A.4: Theoretical and Empirical Values of S_1 and S_2 for each variance vector: (a) Var1, (b) Var2, (c) Var3 when $L = 5 and$a= 2. Standard error values are shown in parentheses.

(a)	Rho	Theoretical		Average	
		S_1	S_2	S_1	S_2
	r1	.047	.057	.044 (.002)	.054 (.001)
	r2	.040	.051	.041 (.002)	.052 (.001)
	r3	-.015	-.019	-.016 (.002)	-.019 (.001)
(b)	r1	.056	.057	.057 (.002)	.057 (.001)
	r2	.042	.051	.043 (.001)	.052 (.001)
	r3	-.029	-.019	-.031 (.001)	-.021 (.001)
(c)	r1	.057	.057	.055 (.001)	.055 (.001)
	r2	.051	.051	.053 (.001)	.053 (.001)
	r3	-.019	-.019	-.016 (.001)	-.016 (.001)

Table A.5: Theoretical and Empirical Values of S_1 and S_2 for each variance vector: (a) Var1, (b) Var2, (c) Var3 when $L = 5 and$a= 4. Standard error values are shown in parentheses.

(a)	Rho	Theoretical		Average	
		S_1	S_2	S_1	S_2
	r1	.102	.143	.101 (.002)	.141 (.001)
	r2	.101	.146	.101 (.002)	.145 (.001)
	r3	-.010	-.046	-.010 (.002)	-.045 (.001)
(b)	r1	.150	.143	.152 (.001)	.143 (.001)
	r2	.136	.146	.136 (.001)	.145 (.001)
	r3	-.074	-.046	-.075 (.001)	-.046 (.001)
(c)	r1	.143	.143	.144 (.001)	.143 (.001)
	r2	.146	.146	.147 (.001)	.146 (.001)
	r3	-.046	-.046	-.046 (.001)	-.046 (.001)

Table A.6: Theoretical and Empirical Values of S_1 and S_2 for each variance vector: (a) Var1, (b) Var2, (c) Var3 when $L = 5 and$a= 6. Standard error values are shown in parentheses.

(a)	Rho	Theoretical		Average	
		S_1	S_2	S_1	S_2
	r1	.182	.240	.183 (.002)	.238 (.001)
	r2	.145	.199	.145 (.002)	.196 (.001)
	r3	-.058	-.081	-.059 (.002)	-.080 (.001)
(b)	r1	.239	.240	.242 (.001)	.241 (.001)
	r2	.179	.199	.179 (.001)	.198 (.001)
	r3	-.108	-.081	-.108 (.001)	-.080 (.001)
(c)	r1	.240	.240	.240 (.001)	.239 (.001)
	r2	.199	.199	.198 (.001)	.197 (.001)
	r3	-.081	-.081	-.085 (.001)	-.085 (.002)

Table A.7: Theoretical and Empirical Values of S_1 and S_2 for each variance vector: (a) Var1, (b) Var2, (c) Var3 when $L = 9 and$a= 2. Standard error values are shown in parentheses.

(a)	Rho	Theoretical		Average	
		S_1	S_2	S_1	S_2
	r1	.059	.072	.059 (.002)	.070 (.001)
	r2	.065	.075	.065 (.002)	.071 (.001)
	r3	.026	.003	.030 (.002)	.004 (.001)
(b)	r1	.081	.072	.077 (.002)	.066 (.001)
	r2	.071	.075	.071 (.002)	.074 (.001)
	r3	-.012	.003	-.014 (.002)	.002 (.001)
(c)	r1	.072	.072	.073 (.001)	.072 (.001)
	r2	.075	.075	.074 (.001)	.074 (.001)
	r3	.003	.003	.005 (.001)	.005 (.001)

Table A.8: Theoretical and Empirical Values of S_1 and S_2 for each variance vector: (a) Var1, (b) Var2, (c) Var3 when $L = 9 and$a= 4. Standard error values are shown in parentheses.

(a)	Rho	Theoretical		Average	
		S_1	S_2	S_1	S_2
	r1	.144	.169	.148 (.002)	.169 (.001)
	r2	.126	.145	.126 (.002)	.145 (.001)
	r3	.050	.071	.051 (.002)	.071 (.002)
(b)	r1	.191	.169	.190 (.001)	.167 (.001)
	r2	.136	.145	.136 (.001)	.143 (.001)
	r3	.063	.071	.063 (.002)	.071 (.002)
(c)	r1	.169	.169	.167 (.001)	.166 (.001)
	r2	.145	.145	.145 (.001)	.144 (.001)
	r3	.071	.071	.070 (.001)	.069 (.002)

Table A.9: Theoretical and Empirical Values of S_1 and S_2 for each variance vector: (a) Var1, (b) Var2, (c) Var3 when $L = 9 and$a= 6. Standard error values are shown in parentheses.

(a)	Rho	Theoretical		Average	
		S_1	S_2	S_1	S_2
	r1	.245	.282	.247 (.002)	.281 (.001)
	r2	.172	.201	.172 (.002)	.197 (.002)
	r3	.025	.037	.022 (.002)	.034 (.002)
(b)	r1	.305	.282	.306 (.001)	.278 (.001)
	r2	.190	.201	.190 (.002)	.199 (.001)
	r3	.049	.037	.051 (.002)	.040 (.002)
(c)	r1	.282	.282	.280 (.001)	.278 (.001)
	r2	.201	.201	.200 (.001)	.199 (.001)
	r3	.037	.037	.036 (.001)	.036 (.002)

Appendix B

Results Tables for Empirical Power Results of S_1 and S_2

This section contains the tables described in Section 5.3.4 detailing the results of the empirical statistical power for each two-sided test procedure under consideration.

Table B.1: Estimates of the Power for S_1 and S_2 for two-sided test of $P_1(|S| \geq c)$ when $L = 3$ with variance vectors: (a) Var1, (b) Var2, (c) Var3 and contiguous alternatives with (i) $a = 2$, (ii) $a = 4$, (iii) $a = 6$

(i)	S_1		S_2		(ii)	S_1		S_2		(iii)	S_1		S_2	
(a)	α				(a)	α				(a)	α			
Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10
r1	.702	.778	.718	.814	r1	1.00	1.00	1.00	1.00	r1	1.00	1.00	1.00	1.00
r2	.322	.462	.322	.438	r2	.958	.972	.940	.972	r2	1.00	1.00	1.00	1.00
r3	.074	.132	.190	.306	r3	.256	.356	.552	.680	r3	.300	.428	.720	.806
	S_1		S_2			S_1		S_2			S_1		S_2	
(b)	α				(b)	α				(b)	α			
Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10
r1	.676	.772	.798	.962	r1	1.00	1.00	1.00	1.00	r1	1.00	1.00	1.00	1.00
r2	.320	.430	.316	.454	r2	.964	.980	.934	.976	r2	1.00	1.00	1.00	1.00
r3	.100	.150	.194	.312	r3	.226	.342	.592	.704	r3	.272	.384	.730	.822
	S_1		S_2			S_1		S_2			S_1		S_2	
(c)	α				(c)	α				(c)	α			
Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10
r1	.700	.814	.694	.814	r1	1.00	1.00	1.00	1.00	r1	1.00	1.00	1.00	1.00
r2	312	.412	.322	.424	r2	.946	.972	.950	.970	r2	1.00	1.00	1.00	1.00
r3	.198	.304	.192	.316	r3	.572	.718	.582	.696	r3	.726	.832	.720	.800

Table B.2: Estimates of the Power for S_1 and S_2 for two-sided test of $P_1(|S| \geq c)$ when $L = 5$ with variance vectors: (a) Var1, (b) Var2, (c) Var3 and contiguous alternatives with (i) $a = 2$, (ii) $a = 4$, (iii) $a = 6$

(i)	S_1		S_2		(ii)	S_1		S_2		(iii)	S_1		S_2	
(a)	α				(a)	α				(a)	α			
Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10
r1	.234	.346	.392	.534	r1	.810	.892	.998	1.00	r1	.998	.998	1.00	1.00
r2	.204	.320	.368	.492	r2	.814	.890	1.00	1.00	r2	.980	.988	1.00	1.00
r3	.076	.130	.096	.162	r3	.056	.108	.438	.564	r3	.402	.524	.708	.794
	S_1		S_2			S_1		S_2			S_1		S_2	
(b)	α				(b)	α				(b)	α			
Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10
r1	.384	.536	.424	.556	r1	.996	.996	.996	.998	r1	1.00	1.00	1.00	1.00
r2	.254	.358	.368	.494	r2	.984	.992	.992	.994	r2	1.00	1.00	1.00	1.00
r3	.154	.240	.090	.188	r3	.628	.748	.316	.436	r3	.894	.936	.710	.804
	S_1		S_2			S_1		S_2			S_1		S_2	
(c)	α				(c)	α				(c)	α			
Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10
r1	.418	.528	.414	.530	r1	.994	.994	.992	.996	r1	1.00	1.00	1.00	1.00
r2	.378	.506	.378	.502	r2	.998	1.00	.998	.998	r2	1.00	1.00	1.00	1.00
r3	.094	.156	.106	.160	r3	.328	.432	.308	.444	r3	.750	.834	.726	.822

Table B.3: Estimates of the Power for S_1 and S_2 for two-sided test of $P_1(|S| \geq c)$ when $L = 9$ with variance vectors: (a) Var1, (b) Var2, (c) Var3 and contiguous alternatives with (i) $a = 2$, (ii) $a = 4$, (iii) $a = 6$

(i)	S_1		S_2		(ii)	S_1		S_2		(iii)	S_1		S_2	
(a)	α				(a)	α				(a)	α			
Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10
r1	.398	.508	.576	.718	r1	.990	.996	1.00	1.00	r1	1.00	1.00	1.00	1.00
r2	.472	.570	.592	.706	r2	.958	.982	.992	.998	r2	.998	1.00	1.00	1.00
r3	.162	.240	.056	.114	r3	.324	.416	.592	.688	r3	.104	.174	.208	.334
	S_1		S_2			S_1		S_2			S_1		S_2	
(b)	α				(b)	α				(b)	α			
Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10
r1	.640	.736	.556	.670	r1	1.00	1.00	1.00	1.00	r1	1.00	1.00	1.00	1.00
r2	.564	.704	.630	.740	r2	.984	.998	.996	.998	r2	1.00	1.00	1.00	1.00
r3	.086	.130	.058	.100	r3	.468	.572	.608	.706	r3	.374	.476	.276	.382
	S_1		S_2			S_1		S_2			S_1		S_2	
(c)	α				(c)	α				(c)	α			
Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10
r1	.640	.732	.638	.734	r1	1.00	1.00	1.00	1.00	r1	1.00	1.00	1.00	1.00
r2	.648	.748	.636	.736	r2	.994	1.00	.994	.996	r2	1.00	1.00	1.00	1.00
r3	.064	.120	.060	.144	r3	.602	.700	.554	.678	r3	.210	.326	.224	.312

Appendix C

Results Tables for Normal Power Results of S_1 and S_2

This section contains the tables described in Section 5.4 detailing the results of power for statistics S_1 and S_2 re-calculated using the normal critical values $z_{\alpha/2}$ = 1.645 and 1.96.

Table C.1: Normal Distribution estimates of Power for S_1 and S_2 for two-sided test of $P_1(|S| \geq c)$ when $L = 3$ with variance vectors: (a) Var1, (b) Var2, (c) Var3 and contiguous alternatives with (i) $a = 2$, (ii) $a = 4$, (iii) $a = 6$

(i)	S_1		S_2		(ii)	S_1		S_2		(iii)	S_1		S_2	
(a)	α				(a)	α				(a)	α			
Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10
r1	.685	.789	.722	.817	r1	1.00	1.00	1.00	1.00	r1	1.00	1.00	1.00	1.00
r2	.338	.463	.328	.450	r2	.962	.981	.942	.970	r2	1.00	1.00	1.00	1.00
r3	.074	.135	.202	.305	r3	.239	.347	.562	.675	r3	.427	.696	.791	.806
	S_1		S_2			S_1		S_2			S_1		S_2	
(b)	α				(b)	α				(b)	α			
Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10
r1	.676	.778	.698	.795	r1	1.00	1.00	1.00	1.00	r1	1.00	1.00	1.00	1.00
r2	.347	.469	.368	.490	r2	.967	.985	.951	.976	r2	1.00	1.00	1.00	1.00
r3	.111	.184	.232	.337	r3	.210	.316	.558	.703	r3	.267	.379	.707	.807
	S_1		S_2			S_1		S_2			S_1		S_2	
(c)	α				(c)	α				(c)	α			
Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10
r1	.729	.821	.726	.818	r1	1.00	1.00	1.00	1.00	r1	1.00	1.00	1.00	1.00
r2	.317	.437	.318	.437	r2	.952	.976	.953	.976	r2	1.00	1.00	1.00	1.00
r3	.183	.278	.184	.279	r3	.610	.720	.600	.708	r3	.717	.818	.711	.807

Table C.2: Normal Distribution estimates of Power for S_1 and S_2 for two-sided test of $P_1(|S| \geq c)$ when $L = 5$ with variance vectors: (a) Var1, (b) Var2, (c) Var3 and contiguous alternatives with (i) $a = 2$, (ii) $a = 4$, (iii) $a = 6$

(i)	S_1		S_2		(ii)	S_1		S_2		(iii)	S_1		S_2	
(a)	α				(a)	α				(a)	α			
Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10
r1	.249	.356	.416	.534	r1	.805	.879	.993	.997	r1	.999	1.00	1.00	1.00
r2	.196	.301	.372	.498	r2	.814	.885	.996	.998	r2	.985	.994	1.00	1.00
r3	.062	.118	.080	.144	r3	.063	.119	.286	.400	r3	.368	.495	.705	.800
	S_1		S_2			S_1		S_2			S_1		S_2	
(b)	α				(b)	α				(b)	α			
Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10
r1	.410	.531	.440	.560	r1	.996	.998	.994	.998	r1	1.00	1.00	1.00	1.00
r2	.313	.429	.422	.547	r2	.984	.993	.996	.999	r2	1.00	1.00	1.00	1.00
r3	.154	.245	.099	.171	r3	.634	.750	.301	.421	r3	.890	.940	.682	.784
	S_1		S_2			S_1		S_2			S_1		S_2	
(c)	α				(c)	α				(c)	α			
Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10
r1	.385	.513	.378	.506	r1	.995	.998	.995	.998	r1	1.00	1.00	1.00	1.00
r2	.370	.494	.374	.498	r2	.997	.999	.997	.999	r2	1.00	1.00	1.00	1.00
r3	.076	.138	.075	.137	r3	.300	.421	.301	.420	r3	.773	.852	.759	.838

Table C.3: Normal Distribution estimates of Power for S_1 and S_2 for two-sided test of $P_1(|S| \geq c)$ when $L = 9$ with variance vectors: (a) Var1, (b) Var2, (c) Var3 and contiguous alternatives with (i) $a = 2$, (ii) $a = 4$, (iii) $a = 6$

(i)	S_1		S_2		(ii)	S_1		S_2		(iii)	S_1		S_2	
(a)	α				(a)	α				(a)	α			
Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10
r1	.373	.497	.570	.688	r1	.984	.993	1.00	1.00	r1	1.00	1.00	1.00	1.00
r2	.443	.564	.619	.728	r2	.956	.978	.995	.998	r2	.998	.999	1.00	1.00
r3	.131	.210	.066	.122	r3	.277	.388	.560	.673	r3	.084	.151	.193	.288
	S_1		S_2			S_1		S_2			S_1		S_2	
(b)	α				(b)	α				(b)	α			
Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10
r1	.672	.771	.598	.706	r1	1.00	1.00	1.00	1.00	r1	1.00	1.00	1.00	1.00
r2	.492	.624	.577	.703	r2	.982	.992	.996	.998	r2	1.00	1.00	1.00	1.00
r3	.079	.141	.063	.119	r3	.443	.570	.583	.695	r3	.329	.446	.247	.350
	S_1		S_2			S_1		S_2			S_1		S_2	
(c)	α				(c)	α				(c)	α			
Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10	Rho	.05	.10	.05	.10
r1	.672	.774	.668	.770	r1	.999	1.00	.999	1.00	r1	1.00	1.00	1.00	1.00
r2	.647	.754	.641	.750	r2	.996	.998	.994	.998	r2	1.00	1.00	1.00	1.00
r3	.061	.116	.060	.115	r3	.614	.721	.598	.704	r3	.192	.290	.213	.310

BIBLIOGRAPHY

[1] Alan Agresti, *Categorical Data Analysis* (Wiley, NJ, 2002)

[2] H. Akaike, "Information theory and an extension of the maximum likelihood principle," In *Second International Symposium on Information Theory*, eds B. N. Petrov and F. Csaki, Akademiai Kiad, Budapest, (1973).

[3] Joshua Albrecht and Rebecca Hwa, "Regression for Sentence-Level MT Evaluation with Pseudo References", In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*,Prague, Czech Republic, (2007).

[4] ALPAC Language and Machines, Computers in Translation and Linguistics. Washington D.C., Publication 1416, National Academy of Sciences, (1966).

[5] M.J. Anderson and C.J.F. ter Braak, "Permutation Tests for Multi-factorial Snalysis of Variance", Journal of Statistical Computation and Simulation, **73**, (2003).

[6] D.J. Arnold, Lorna Balkan, Siety Meijer, R.Lee Humphreys and Louisa Sadler, *Machine Translation: An Introductory Guide* (Blackwells-NCC, London, 1994).

[7] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," In *Proceedings*

of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL '05), Ann Arbor, Michigan, (2005).

[8] George E.P. Box, William G. Hunter, and J. Stuart Hunter, *Statistics for Experimenters* (Wiley, New York, 1978).

[9] B. Broome, A. Brodeen, F. Brundick and M. Taylor. "A Quantitative Method for Evaluating Machine Translation Systems," In *Proceedings of the Sixth Annual Army Conference on Applied Statistics* (ACAS '00), Houston, TX, (2000).

[10] Jean Carletta, "Assessing Agreement on Classification Tasks: The Kappa Statistic," Computational Lingusitics,**22**, 2, (1996).

[11] George Casella and Roger L. Berger, *Statistical Inference* (Duxbury, Australia, 2002).

[12] K. Church and E. Hovy, "Good Applications for Crummy Machine Translation," Machine Translation, **8**, (1993).

[13] William S. Cleveland, "Robust Locally Weighted Regression and Smoothing Scatterplots," Journal of the American Statistical Association, **74**, 368, (1979).

[14] W. J. Conover, *Practical Nonparametric Statistics* (Wiley, New York, 1980).

[15] W.J. Conover and Ronald L. Iman, "Rank Transformations as a Bridge Between Parametric and Nonparametric Statistics," The American Statistician, **35**, 3, (1981).

[16] Deborah Coughlin, "Correlating Automated and Human Assessments of Machine Translation Quality," Proceedings of MT Summitt IX,(2003).

[17] Damianos, Laurie User-Centered Evaluation. Briefing, MITRE Corporation. http://www.mitre.org/work /tech papers/tech papers 01/damianos evaluation/ damianos evaluation.pdf (2001).

[18] Annette J. Dobson, *An Introduction to Generalized Linear Models* (Chapman & Hall, London, 1990.

[19] Doddington, George "Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics." (2002). In P̂roceedings of the Second Conference on Human Language Technology San Diego, CA pp. 128-132.(HLT 2002).

[20] Pamela W. Jordan, Bonnie J. Dorr, and John W. Benoit, "A First Pass Approach for Evaluating Machine Translation Systems", Machine Translation, **8**, 1993.

[21] J. Doyon, K. Taylor, and J. White, "Task-based Evaluation of Machine Translation", In *Proceedings of Machine Translation Summit VII*, Singapore, (1999).

[22] Ertoz, L., M. Steinback, and V. Kumar. A New Shared Nearest Neighbor Clustering Algorithm and its Application. In Proceedings of the workshop on Clustering High Dimensional Data and its Applications. (2002).

[23] Fisher, F., C. Schlesiger, L. Decrozant, R.Zuba, M. Holland, and C.R. Voss. Searching and Translating Arabic Documents on a Mobile Platform. Proceed-

ings of the Advanced Information Processing and Analysis Conference (AIPA 99) . McLean, VA. (1999).

[24] John Fox *Applied Regression Analysis, Linear Models, and Related Methods* (Sage, California, 1997).

[25] Jess Gimnez, Enrike Amig, and Chiori Hori, "Machine Translation Evaluation Inside QARLA", In *Proceedings of the International Workshop on Spoken Language Translation: Evaluation Campaign on Spoken Language Translation* (IWSLT 2005), Pittsburgh, PA (2005).

[26] P. Good, *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses* (Springer, New York,1994).

[27] Phillip I. Good, *Resampling Methods: A Practical Guide to Data Analysis* (Birkhäuser, Boston, 1999).

[28] Trevor Hastie and Clive Loader "Local Regression: Automatic Kernel Carpentry", Statistical Science, **8**, 81-93. (1993).

[29] Rod Holland. Embedded Machine Translation Prototypes at MITRE. Presentation at UMIACS Computational Linguistics Colloquium.University of Maryland, College Park. (2005).

[30] David W. Hosmer and Stanley Lemeshow , *Applied Logistic Regression* (Wiley, New York, 2000).

[31] Hovy, E. Toward Finely Differentiated Evaluation Metrics for Machine Translation. Proceedings of the EAGLES Workshop on Standards and Evaluation. Pisa, Italy. (1999).

[32] International Standards for Language Engineering. (http://www.isi.edu/natural-language/mteval) The ISLE Classification of Machine Translation Evaluations, Draft 1, October, 2000. Proceedings of the Hands-on Workshop on Machine Translation Evaluation. Association for Machine Translation in the Americas, Cuernavaca, Mexico. (2000).

[33] Doug Jones, Wade Shen, Neil Granoien, Martha Herzog, and Clifford Weinstein. "Measuring Translation Quality by Testing English Speakers with a New Defense Language Proficiency Test for Arabic". ıInternational Conference on Intelligence Analysis. McLean, VA. (2005).

[34] M. Kendall, "Partial Rank Correlation", Biometrika, **30**, 81-93. (1942).

[35] Margaret King, Andrei Popescu-Belis and Eduard Hovy. 2003. FEMTI: Creating and Using a Framework for MT Evaluation. In Proceedings of MT Summit IX, New Orleans, LA. pp. 224-231.(2003).

[36] Samuel Kotz, Norman Lloyd Johnson, and Campbell B. Read (Eds.). *Encyclopedia of Statistical Sciences*

[37] Grazia Russo-Lassner, Jimmy Lin, and Philip Resnik, "A Paraphrase-Based Approach to Machine Translation Evaluation", Technical Report UMIACS-TR-2005-57, University of Maryland, College Park, (2005)

[38] Alon Lavie, Kenji Sagae and Shyamsundar Jayaraman, "The Significance of Recall in Automatic Metrics for MT Evaluation". In Proceedings of AMTA-2004, Washington DC. (2004).

[39] A. Lavie and A. Agarwal. "METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments", Proceedings of Workshop on Statistical Machine Translation at the 45th Annual Meeting of the Association of Computational Linguistics (ACL-2007), Prague, June 2007

[40] Nicole Lazar and Per Aslak Mykland, "An Evaluation of the Power and Conditionality Properties of Empirical Likelihood", Biometrika,**85**, 3, (1998).

[41] Lehmann, E. L. and D'Abrera, H. J. M. Nonparametrics: Statistical Methods Based on Ranks, rev. ed. Englewood Cliffs, NJ: Prentice-Hall, pp. 292, 300, and 323, 1998.

[42] Gregor Leusch, Nicola Ueffing and Hermann Ney. "CDER: Efficient MT Evaluation Using Block Movements." Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006).

[43] L. Levin, B. Bartlog, A. Llitjos, D. Gates, A. Lavie, D. Wallace, T. Wantanabe and M. Woszczyna, "Lessons Learned from a Task-based evaluation of speech to speech machine translation", In *Proceedings of the Language Resources Conference* (LREC), Athens, Greece, (2000).

[44] Chin-Yew Lin and Franz Josef Och, "Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics," In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics* (ACL '04), Barcelona, Spain, (2004).

[45] Ding Liu and Daniel Gildea, "Source-language Features and Maximum Correlation Training for Machine Translation Evaluation", In *Proceedings of the HLT/NAACL-2007*, (2007).

[46] Robert L. Mason, Richard G. Gunst, and James L. Hess, *Statistical Design and Analysis of Experiments* (Wiley, New York, 1989).

[47] P. McCullagh and J.A. Nelder, *Generalized Linear Models* (Chapman & Hall, London, 1989).

[48] Charles E. McCulloch and Shayle R. Searle, *Generalized, Linear, and Mixed Models* (John Wiley & Sons, New York, 2001).

[49] I. Dan Melamed "Automatic Evaluation and Uniform Filter Cascades for Inducing N-Best Translation Lexicons," In *Proceedings of Third Workshop on Very Large Corpora* (WVLC3), Boston, Massachusetts, (1995).

[50] I. Dan Melamed, Ryan Green and Joseph P. Turian,"Precision and Recall of Machine Translation," In *Proceedings of the Human Language Technology Conference* (HLT-NAACL '03), Edmonton, Canada, (2003).

[51] Joseph Olive, Global Autonomous Language Exploitation (GALE), DARPA/IPTO Proposer Information Pamphlet, (2005).

[52] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (ACL '02), Philadelphia, Pennsylvania, (2002).

[53] R Development Core Team, R: A language and environment for statistical computing, ıR Foundation for Statistical Computing, http://www.R-project.org , Vienna, Austria. (2004).

[54] Florence Reeder and John White, "Granularity in MT Evaluation," In *Proceedings of Towards Systematizing MT Evaluation: A Workshop on Machine Translation Evaluation at the MT Summit IX*, New Orleans, LA, (2003).

[55] Philip Resnik, "Evaluating Multilingual Gisting of Web Pages," In *Proceedings of the AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*," Stanford University, (1997).

[56] J. C. Sager, "Quality and Standards: The Evaluation of Translations," In *The Translators Handbook* C. Picken (Eds.), London, (1989).

[57] Thomas A. Severini, *Elements of Distribution Theory* (Cambridge University Press, New York, 2005).

[58] M. Snover, B. J. Dorr, R. Schwartz, J. Makhoul, L. Micciulla, and R. Weischedel, "A Study of Translation Error Rate with Targeted Human Annotation," Technical Report LAMP-TR-126, CS-TR-4755, UMIACS-TR-2005-58, University of Maryland, College Park, (2005).

[59] Karen Spärck-Jones and J.R. Gallier, *Evaluating Natural Language Processing Systems: An Analysis and Review* (Springer, Berlin, 1996).

[60] Calandra R. Tate and Clare R. Voss,"Combining Evaluation Metrics Via Loss Functions," In *Proceedings of the Association for Machine Translation in the Americas* (AMTA '06), Boston, MA, (2006).

[61] Kathryn Taylor and John White, "Predicting What MT is Good for: User Judgements and Task Performance," In Proceedings of the Third Conference of the Association for Machine Translation in the Americas (AMTA '98), Langhorne, PA, (1998).

[62] Joseph P. Turian, Luke Shen and I. Dan Melamed. "Evaluation of Machine Translation and its Evaluation," In *Proceedings of MT Summit IX*, New Orleans, LA, (2003).

[63] M. Vanni and K. Miller, "Scaling the ISLE Framework: Validating Tests of Machine Translation Quality for Multi-Dimensional Measurement," In *Proceedings of the Workshop on MT Evaluation at the MT Summit VIII*, Santiago de Campostela, Spain, (2001).

[64] M. Vanni and K. Miller, "Scaling the ISLE Framework: Use of Existing Corpus Resources for Validation of MT Evaluation Metrics Across Languages," In *Proceedings of the Language Resources and Evaluation Conference* (LREC '02), Las Palmas, Canary Islands, Spain., (2002).

[65] M. Vanni, C. Voss, and C. Tate, "Ground Truth, Reference Truth & Omniscient Truth - Parallel Phrases in Parallel Texts for MT Evaluation," In *Proceedings of LREC*, Lisbon, Portugal, (2004).

[66] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S-Plus*, (Springer, New York, 1999).

[67] C. R. Voss. MT Evaluation: Measures of Effectiveness in Document Exploitation. DARPA TIDES PI Meeting. Santa Monica, CA. (2002).

[68] Clare R. Voss, Calandra R. Tate, and Eric Slud. Task-based Machine Translation Evaluation. ıPresentation at MT Evaluation Panel, AMTA. Georgetown University, Washington, DC. (2004).

[69] Clare R. Voss and Calandra R. Tate, "Task-based Evaluation of Machien translation (MT) Engines: Measuring How Well People Extract Who, When, Where-Type Elements in MT Output", In Proceedings of European Association for Machine Translation Conference (EAMT), Oslo, Norwar, (2006).

[70] J. White and T. O' Connell, "The ARPA MT Evaluation Methodologies: evolution, lessons, and future approaches", In *Proceedings of the Association for Machine Translation in the Americas Conference* (AMTA), (1994).

[71] Y Wilks, "Keynote: Traditions in Evaluation of MT", In *Proceedings of a workshop sponsored by the National Science Foundation*, San Diego, CA, (1994).